GUILHERME PUMI
TAIANE SCHAEDLER PRASS
(ORGANIZADORES)

APLICAÇÕES DE MACHINE LEARNING:

Resultados de um Curso de Verão

GUILHERME PUMI
TAIANE SCHAEDLER PRASS
(ORGANIZADORES)

APLICAÇÕES DE MACHINE LEARNING:

Resultados de um Curso de Verão

CIÊNCIAS EXATAS E DA TERRA E ENGENHARIAS

Aplicações de machine learning: resultados de um curso de verão

Diagramação: Ellen Andressa Kubisty
Correção: Yaiddy Paola Martinez
Indexação: Amanda Kelly da Costa Veiga
Revisão: Os autores
Organizadores: Guilherme Pumi
Taiane Schaedler Prass

Dados Internacionais de Catalogação na Publicação (CIP)

A642 Aplicações de machine learning: resultados de um curso de verão / Organizadores Guilherme Pumi, Taiane Schaedler Prass. – Ponta Grossa - PR: Atena, 2024.

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader
Modo de acesso: World Wide Web
Inclui bibliografia
ISBN 978-65-258-2439-0
DOI: https://doi.org/10.22533/at.ed.390241804

1. Inteligência Artificial - Aprendizado de Máquina. I. Pumi, Guilherme (Organizador). II. Prass, Taiane Schaedler (Organizadora). III. Título.

CDD 006.3

Elaborado por Bibliotecária Janaina Ramos – CRB-8/9166

Atena Editora
Ponta Grossa – Paraná – Brasil
Telefone: +55 (42) 3323-5493
www.atenaeditora.com.br
contato@atenaeditora.com.br

DECLARAÇÃO DOS AUTORES

Os autores desta obra: 1. Atestam não possuir qualquer interesse comercial que constitua um conflito de interesses em relação ao artigo científico publicado; 2. Declaram que participaram ativamente da construção dos respectivos manuscritos, preferencialmente na: a) Concepção do estudo, e/ou aquisição de dados, e/ou análise e interpretação de dados; b) Elaboração do artigo ou revisão com vistas a tornar o material intelectualmente relevante; c) Aprovação final do manuscrito para submissão.; 3. Certificam que os artigos científicos publicados estão completamente isentos de dados e/ou resultados fraudulentos; 4. Confirmam a citação e a referência correta de todos os dados e de interpretações de dados de outras pesquisas; 5. Reconhecem terem informado todas as fontes de financiamento recebidas para a consecução da pesquisa; 6. Autorizam a edição da obra, que incluem os registros de ficha catalográfica, ISBN, DOI e demais indexadores, projeto visual e criação de capa, diagramação de miolo, assim como lançamento e divulgação da mesma conforme critérios da Atena Editora.

DECLARAÇÃO DA EDITORA

A Atena Editora declara, para os devidos fins de direito, que: 1. A presente publicação constitui apenas transferência temporária dos direitos autorais, direito sobre a publicação, inclusive não constitui responsabilidade solidária na criação dos manuscritos publicados, nos termos previstos na Lei sobre direitos autorais (Lei 9610/98), no art. 184 do Código Penal e no art. 927 do Código Civil; 2. Autoriza e incentiva os autores a assinarem contratos com repositórios institucionais, com fins exclusivos de divulgação da obra, desde que com o devido reconhecimento de autoria e edição e sem qualquer finalidade comercial; 3. Todos os e-book são *open access, desta forma* não os comercializa em seu site, sites parceiros, plataformas de *e-commerce,* ou qualquer outro meio virtual ou físico, portanto, está isenta de repasses de direitos autorais aos autores; 4. Todos os membros do conselho editorial são doutores e vinculados a instituições de ensino superior públicas, conforme recomendação da CAPES para obtenção do Qualis livro; 5. Não cede, comercializa ou autoriza a utilização dos nomes e e-mails dos autores, bem como nenhum outro dado dos mesmos, para qualquer finalidade que não o escopo da divulgação desta obra.

Havia tempos que planejávamos propor uma disciplina para o Programa de Pós-Graduação em Estatística (PPGEst-UFRGS), ao nível de mestrado, tratando de aspectos estatísticos associadas a técnicas de machine learning. Tínhamos a ideia de um curso orientado para a discussão dos aspectos teóricos de machine learning, da teoria assintótica por trás desses modelos e de suas intersecções e ramificações com a estatística.

Durante o ano de 2019, os Programas de Pós-Graduação em Matemática e Matemática Aplicada (PPGMat e PPGMap) da UFRGS planejavam uma segunda edição de seu programa de verão, que havia ocorrido nos meses de janeiro e fevereiro daquele ano no Instituto de Matemática e Estatística (IME) da UFRGS, com grande sucesso. Contando com o aval do PPGEst, propomos uma parceria entre o PPGEst, o PPGMat e o PPGMap para a realização do programa do verão. Vimos nesta parceria a possibilidade de ofertar a disciplina de machine learning que planejávamos. Após alguma discussão e muita preparação nascia a disciplina de Machine Learning e Modelagem Estatística que idealizamos.

A procura pelo curso foi bastante forte, motivada talvez pelo assunto, talvez pela pouca oferta de programas de verão voltados à estatística naquele ano. Após um difícil processo, dada a procura altamente qualificada que recebemos, selecionamos 35 candidatos das mais variadas áreas, como estatística, matemática, física, ciências da computação, engenharia, economia, química, entre outras e também com a mais variada formação, como graduandos, mestrandos, doutorandos e profissionais do mercado. No início de 2020 iniciávamos o curso de verão em meio à notícias vindas da China sobre uma variante do vírus da COVID que estava causando preocupações. Felizmente, o curso procedeu sem incidentes. Porém, poucas semanas após terminado o curso de verão, a COVID 19 chegava com toda a força ao Brasil...

A avaliação do curso foi um trabalho proposto para ser feito em grupos com temática livre, onde os participantes deveriam escolher e analisar dados disponíveis livremente na internet, utilizando alguma das técnicas aprendidas durante o curso. As avaliações se dariam através de um relatório em formato de artigo e de uma apresentação. O resultado das avaliações foi tão bom que decidimos conjuntamente com os alunos, editar um livro com os artigos, após a devida formatação, correções e sugestões dadas por nós.

Assim nasceu o presente livro. Os artigos aqui presentes representam de forma muito fiel a qualidade dos trabalhos apresentados no curso. Frisamos que os artigos foram apenas revisados por nós, não tendo passado por uma revisão por pares como normalmente seriam, caso submetidos para publicação em revistas especializadas. A maioria das mudanças, correções e sugestões feitas por nós foram relacionadas à formatação, apresentação e linguagem.

Agradecemos imensamente a todos os participantes pelo empenho e dedicação ao curso, ao PPGEst, à direção do Instituto de Matemática e Estatística da UFRGS e ao PPGMat e PPGMap pela parceria.

Guilherme Pumi e Taiane Schaedler Prass
Porto Alegre, Agosto de 2021

SUMÁRIO

SUMÁRIO

IDENTIFICAÇÃO DO PERFIL SOCIOPSICOLÓGICO DE USUÁRIOS DE DROGAS

Albertine Weber Carneiro†
Instituto de Física - UFRGS

Eduardo Gressler Brock
Programa de Pós-Graduação em Física - UFRGS

Mayara Bello Soares
Instituto de Matemática e Estatística - UFRGS

Taís Loureiro Bellini
Programa de Pós-Graduação em Estatística - UFRGS

RESUMO: Este trabalho tem como objetivo avaliar os principais fatores que identificam o perfil sociopsicológico de usuários de drogas. Para tal, realizou-se a análise de problemas de classificação quanto ao uso de quatro drogas: álcool, cannabis, ecstasy e um grupo de drogas estimulantes (cocaína, crack e anfetamina). O uso dessas drogas foi avaliado em relação a características sociais e psicológicas dos indivíduos por meio do uso de dois algoritmos de machine learning: regressão logística e árvores de decisão. Os resultados obtidos permitem traçar um perfil para o usuário de drogas com acurácia e sensibilidade consideradas satisfatórias.
PALAVRAS-CHAVE: Uso de Drogas, Machine Learning, Classificação, Regressão Logística e Árvore de Decisão.

1.1 INTRODUÇÃO

Avaliar o potencial risco do consumo individual de drogas, ilícitas ou não, é um problema de saúde pública: além dos danos que podem ser causados à saúde pelo seu simples uso, as drogas (em particular, as ilícitas) podem implicar em severas consequências para a sociedade, como o surgimento de redes de tráfico ou sujeitando os seus usuários a viverem em condiçoes vulneráveis. Embora muitas vezes as consequências possam ser potencializadas ao indivíduo em questão, o uso de drogas é entendido como um fenômeno não isolado e frequentemente relacionado a certos grupos comportamentais pelo senso comum (Fehrman et al., 2015b). Por conta desse fato, muitos estudos tentam abordar o consumo de drogas de forma a identificar quais são esses grupos de risco, de maneira a desenvolver políticas públicas mais direcionadas a esses indivíduos.

Motivados por esse panorama e por considerarmos essa questão relevante para sociedade atual, a modelagem do uso de drogas foi escolhida pelo grupo como temática para o presente trabalho. Nele, iremos abordar diferentes técnicas estatísticas de classificação

com o objetivo de encontrar características demográficas e de personalidade dos indivíduos que estejam associadas ao uso de drogas. Para tal, será utilizado um conjunto de dados disponível em Fehrman et al. (2015a) resultantes de uma pesquisa realizada online sobre o tema, que serviu de base para um extenso trabalho com ênfase na área da psicologia acerca do tema (Fehrman et al., 2015b).

Serão avaliadas apenas três drogas do conjunto relacionado pela pesquisa, além de um grupo de drogas específicas. As drogas foram escolhidas representando drogas lícitas e ilícitas, sendo elas: álcool, maconha (cannabis), ecstasy e um grupo de drogas estimulantes: cocaína, crack e anfetamina. Para a separação dos grupos, foi observado o desfecho de usúario e não usúario de cada droga, considerando usuários aqueles que fizeram uso da determinada substância em um período inferior a um ano até o momento da realização da pesquisa.

As análises foram feitas separadamente para cada droga e para o grupo, dado que as distribuições nos atributos estudados pela pesquisa diferem consideravelmente e supondo que diferentes perfis comportamentais podem ser encontrados dependendo da droga avaliada. As técnicas aplicadas foram regressão logística e Árvore de decisão, obtendo resultados satisfatórios na faixa dos 70% de acurácia na classificação. Os modelos resultantes foram avaliados principalmente de acordo com os critérios de acurácia e de sensibilidade: enquanto o primeiro mede a taxa de acerto do ajuste, o segundo mede a capacidade do algoritmo de identificar corretamente os usuários de drogas, que são o foco do trabalho.

1.2 DESCRIÇÃO DOS DADOS

A base de dados foi coletada a partir de uma pesquisa anônima online, realizada entre março de 2011 e março de 2012. A amostra foi obtida utilizando a amostragem por bola de neve e não foi possível identificar se havia um público alvo definido previamente pela pesquisa. Foram obtidas 2051 respostas, das quais 166 foram retiradas do banco por estarem mal preenchidas, totalizando 1885 observações restantes. A base de dados é constituída por 32 variáveis, sendo a primeira, *ID*, apenas a ordenação da resposta na pesquisa. Dentre as restantes, 12 estão associadas com fatores sociais e comportamentais. São elas: idade, gênero, educação, país, neuroticismo, extroversão, abertura, socialização, conscienciosidade, impulsividade e busca por sensações.

As demais 19 variáveis são ligadas ao uso das diferentes substâncias consideradas no estudo, a saber: **Álcool**, **Anfetamina**, Nitritos de Alquila, Benzodiazepina, Cafeína, **Cannabis**, Chocolate, **Cocaína**, **Crack**, **Ecstasy**, Heroína, Ketamina, Legal Highs, LSD, Metadona, Cogumelos, Nicotina, Semeron e Inalantes (VSA), onde as substâncias em destaque são as variáveis utilizadas no presente estudo. Todos os respondentes classificaram o uso de cada uma das substâncias listadas acima em 7 possíveis categorias,

de acordo com a frequência de seu uso: nunca usou, usou há mais de uma década, usou na última década, no último ano, no último mês, na última semana ou no último dia. Uma característica importante de ser ressaltada sobre essa pesquisa é que, devido a forma como a pesquisa foi conduzida, o banco de dados gerado mostrou-se extremamente viesado. Em sua maioria, as respostas foram de pessoas de nível educacional elevado, brancas e residentes ou do Reino Unido ou dos EUA, e a proporção de usuários de drogas ilegais na amostra é bem mais elevada do que de fato existe entre as populações gerais.

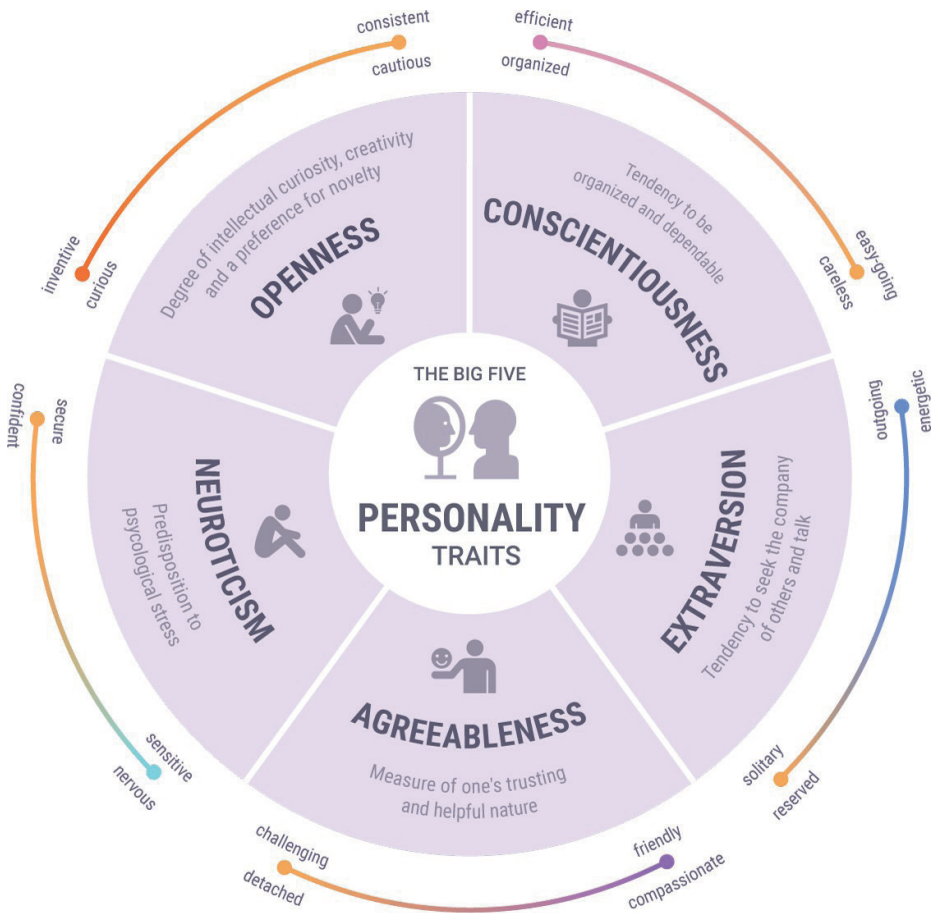
1.2.1 Traços de Personalidade (NEO-FFI-R)

A pesquisa utilizou o teste de traços de personalidade NEO-FFI-R, um questionário composto por 60 itens pontuados de 1 (*discordo fortemente*) a 5 (*concordo fortemente*). O questionário permite avaliar as cinco dimensões básicas da personalidade através do modelo de cinco grandes fatores (*Big Five Factors* - modelo amplamente utilizado pela comunidade da área da psicologia desde a década de 1930): neuroticismo, extroversão, abertura, socialização, conscienciosidade. A Figura 1.1 apresenta uma representação esquemática dos cinco fatores, juntamente de alguns comportamentos previstos para pessoas com pontuações altas e baixas associadas a cada fator.

Abaixo uma breve descrição de cada traço de personalidade definido no modelo de cinco grandes fatores, conforme Faria (2014):

- **Neuroticismo (N)** - Mede a instabilidade emocional. Pessoas com pontuações altas nessa escala tendem a ser ansiosas, inibidas, melancólicas e dotadas de baixa autoestima. Já as que obtém baixa pontuação tendem a ser de fácil trato, otimistas e dotadas de boa estima consigo mesmas;
- **Extroversão (E)** - É a mais ampla das cinco dimensões. Mede a sensação de bem-estar, o nível de energia e a habilidade nas relações interpessoais. Pontuações elevadas significam afabilidade, sociabilidade e capacidade de se impor. Baixas indicam introversão, reserva e submissão;
- **Abertura (O)** - Pessoas com pontuações elevadas gostam de novidades e tendem a ser criativas. Na outra ponta da escala estão os convencionais e ordeiros, os que gostam da rotina e têm senso aguçado do certo e do errado;
- **Socialização (A)** - Refere-se ao modo como nos relacionamos com os outros. Muitos pontos indicam uma pessoa compassiva, amistosa e calorosa. Na outra extremidade estão os retraídos, críticos e egocêntricos;
- **Conscienciosidade (C)** - Mede o grau de concentração. Aqueles com altas pontuações apresentam grande motivação, são disciplinados, comprometidos e confiáveis. Os que apresentam resultados baixos são indisciplinados e se distraem facilmente.

Apesar dos scores possuírem resultados discretos, eles com frequência são agrupados em categorias: scores abaixo de 35 são categorizados como "Very Low", entre 35 e 45 eles são considerados "Low", entre 45 e 55 temos os scores "Average", entre 55 e 65 temos "High" e acima de 65 possuímos scores na faixa "Very High".

Figura 1.1: Os cinco grandes fatores de personalidade.

Fonte: J.M. Digman, "personality structure emergence of the five factor model"

1.2.2 Alterações nos dados

Os atributos que podem ser utilizados para classificar o uso de drogas são os 12 fatores sociais e comportamentais listados anteriormente. Desses 12, dois atributos, impulsividade e busca por sensações, foram desconsiderados pela falta de embasamento teórico acerca do assunto pelos membros do grupo e de referências confiáveis quanto à classificação dos mesmos. Dessa forma, restam 10 atributos que podem ser divididos em 2 grupos: o grupo de indicadores sociais, composto por Idade, Gênero, Educação, País e Etnia, e o grupo de indicadores comportamentais, composto pelos 5 Scores do NEO-FFI-R (N, E, O, A, C). Para ambos os grupos, analisamos a distribuição de cada variável entre seus diferentes níveis, para decidir se todas as variáveis entrariam na análise e se era necessário alterar algumas categorias.

Indicadores Sociais

Os resultados da análise de distribuição para os Indicadores Sociais podem ser observados na Figura 1.2. A partir dessa análise, foi possível observar que dois indicadores sociais apresentavam uma concentração excessiva em categorias específicas: País concentra-se em torno de dois níveis (UK 55% e USA 30%), enquanto Etnia está totalmente concentrada no nível “White” (População Branca tem 91% contra apenas 9% dos demais). Deste modo, optou-se por retirar a priori ambas as variáveis das análises realizadas, a fim de não prejudicar os resultados obtidos.

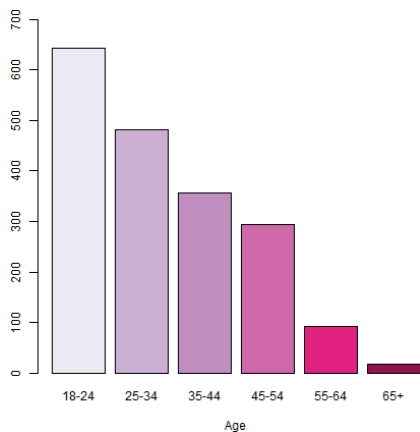
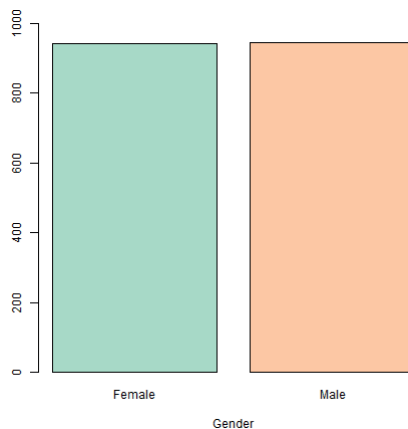
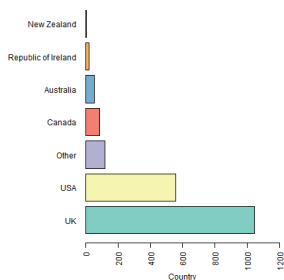
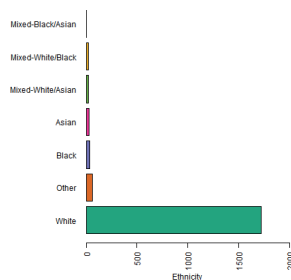
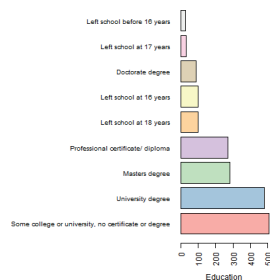
Dos 3 indicadores sociais remanescentes, apenas gênero possui uma distribuição bem balanceada. Para Idade, há uma concentração de indivíduos na faixa de 18 a 24 anos e 25 a 34 anos, compondo mais de 50% da amostra. Na variável educação, observa-se a moda de indivíduos na categoria de estudantes sem diploma, com maior concentração de respondentes nas classe de cursos profissionalizantes a diploma de mestrado. Para podermos lidar melhor com essas variáveis, optou-se por agrupar algumas categorias: em Idade, as categorias “45-54”, “55-64” e “65+” foram agrupadas de maneira a criarmos a categoria “45+”. Para educação, as categorias “Left School before 16 years”, “Left School at 16 years”, “Left School at 17 years” e “Left School at 18 years” foram todas agrupadas para criar a categoria “Left School”. Além disso, as categorias “Doctorate degree” e “Masters degree” foram agrupadas na categoria “Graduate degree”. As demais categorias se mantiveram inalteradas.

(a) Idade (Age)

(b) Gênero (Gender)

(c) País (Country)

(d) Etnia (Ethnicity)

(e) Educação (Education)

Figura 1.2: Distribuição dos Indicadores Sociais: (a) idade, (b) gênero, (c) país, (d) etnia e (e) educação.

Indicadores Comportamentais

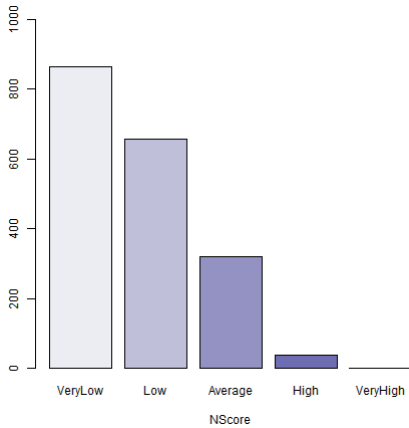
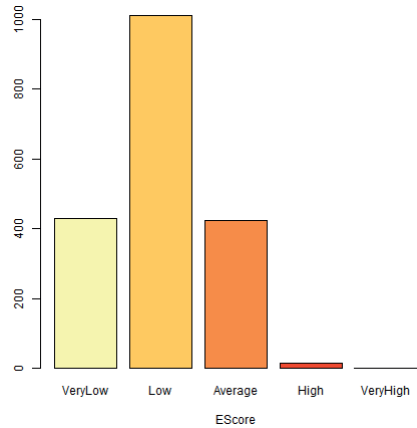
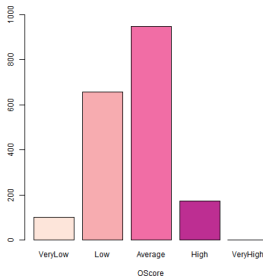
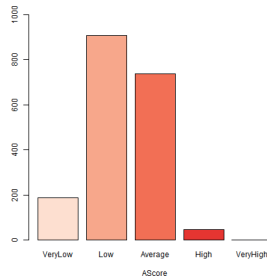
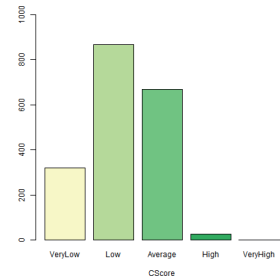
Os 5 indicadores comportamentais foram agrupados nas categorias de “Very Low” até “Very High” definidas anteriormente na Seção 1.2.1 e tiveram suas distribuições observadas. Apesar de eles também apresentarem concentração em torno de algumas categorias específicas, optou-se por não aplicar nenhuma transformação sobre eles, devido ao fato de serem indicadores psicológicos pré-definidos e o grupo não possuir conhecimento suficiente da área para realizar outros agrupamentos. As distribuições de cada Score podem ser observadas na Figura 1.3.

(a) Neuroticismo (b) Extroversão

(c) Abertura (d) Socialização (e) Conscienciosidade

Figura 1.3: Distribuição dos Indicadores Comportamentais: (a) Neuriticismo, (b) extroversão, (c) abertura, (d) socialização e (e) conscienciosidade.

Exclusão de Observações

Além das transformações nas variáveis, foram excluídas 8 entradas pertencentes a usuários que responderam que já haviam utilizado em qualquer momento “Semeron”, uma droga fictícia incluída no estudo para identificar participantes da pesquisa cujas respostas não fossem confiáveis.

Definição das Variáveis Respostas

Por fim, a última transformação na base de dados foi realizada em relação às drogas: ao invés de trabalhar com as 19 drogas listadas e com as 7 categorias diferentes de cada uma delas, optou-se por escolher subgrupos para a análise. Em primeiro lugar, decidiu-se trabalhar com um problema de classificação binária - ou seja, apenas classificar se o indivíduo é usúario ou não-usúario de uma certa droga. Para tal, as categorias de “nunca usou”, “usou há mais de uma década” e “usou na última década” foram consideradas como “Não Usuários”, enquanto as categorias “usou no último ano”, “usou no último mês”,

“usou na última semana” e “usou no último dia” foram consideradas como “Usuários”. Essa escolha foi feita por considerarmos que a utilização de drogas no período de um ano é uma janela razoável para considerar um indivíduo como usuário, enquanto o período de uma década (que seria o próximo agrupamento possível) já é amplo demais.

Em uma segunda etapa, escolheu-se subconjuntos das 19 drogas para serem analisados. Para tal, realizamos a escolha das drogas baseada em dois critérios: presença razoável de usuários da droga na base (já considerando o período de 1 ano) e popularidade da droga no Brasil (alguma das drogas listadas na pesquisa são mais disseminadas no exterior, sendo pouco conhecidas no Brasil). Dessa forma, foram definidos 4 problemas de classificação diferentes: um envolvendo o uso de uma droga lícita - álcool, um para uma droga ilícita com distribuição bem balanceada - maconha, um para uma droga ilícita com uma distribuição não tão balanceada (70% - 30%) - ecstasy e um para um conjunto de drogas agrupadas por possuírem efeito estimulante - cocaína, crack e anfetamina. A distribuição entre usuários e não usuários para cada um desses 4 problemas podem ser observadas na Figura 1.4.

(a) Álcool

(b) Maconha (Cannabis)

(c) Ecstasy

(d) Drogas Estimulantes

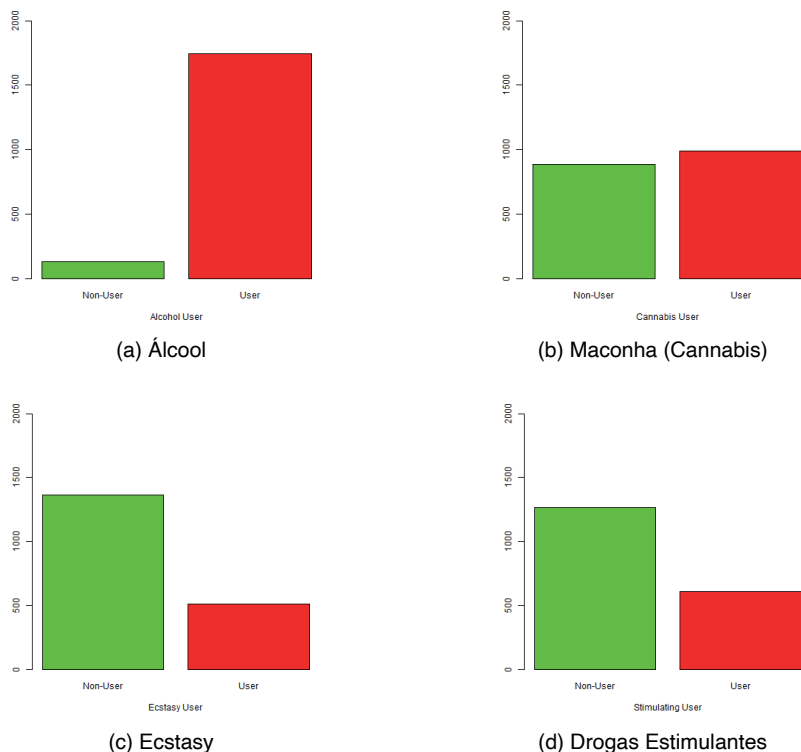



Figura 1.4: Distribuição dos usuários de drogas. (a) álcool, (b) maconha, (c) ecstasy e (d) drogas estimulantes.

1.2.3 Resumo dos dados

Ao final de todo esse processo de transformações, obtivemos uma base de 1877 observações de 12 variáveis, sendo 8 delas atributos (3 indicadores sociais - idade, gênero, educação - e 5 indicadores comportamentais - neuroticismo, extroversão, abertura, socialização e conscienciosidade) e 4 respostas sobre o uso de diferentes drogas - álcool, maconha (cannabis), ecstasy e o grupo de drogas estimulantes, composto por cocaína, crack e anfetamina.

No Apêndice são apresentadas as Tabelas 1.38 e 1.39 que resumem as informações a respeito dos respondentes da pesquisa. Nessas tabelas são descritas as frequências e proporções de cada classe considerando o desfecho para cada droga ou grupo de drogas, já considerando todas as transformações realizadas.

1.3 MÉTODOS UTILIZADOS

O objetivo desde estudo é criar um classificador que, baseado nas variáveis descritas na Seção 1.2, identifique a qual grupo um determinado indivíduo pertence, se usuário ou não de determinada droga ou do grupo de drogas.

Foram avaliados dois métodos de classificação: regressão logística e Árvore de decisão. Em cada método, para cada droga ou grupo de drogas, foi avaliada a possibilidade de reduzir o número de variáveis independentes de forma a manter ou melhorar o resultado. Os modelos finais de cada método serão apresentados e analisados, visando identificar perfis de usuários. Ao final, serão comparados os métodos para definir qual será melhor de acordo com métricas de acurácia e sensibilidade padrão.

Para ambos os métodos, dividimos a amostra em treinamento (85% das observações) e teste (15% das observações) utilizando a função **createDataPartition** do pacote **caret** do R (versão 3.6.2) para garantir distribuições similares. A semente utilizada nessa separação foi 205650. Os códigos detalhados de todos os métodos aplicados podem ser conferidos no repositório do Github utilizado pelo grupo (Bellini et al., 2020).

1.3.1 Regressão Logística

Considerando o interesse em avaliar o desfecho dicotômico de usuários e não usuários, e buscando entender a relação entre a variável resposta e as demais variáveis independentes, optou-se por utilizar regressão logística. Este é um método comumente aplicado a esse tipo de modelagem (Araujo and de Montreuil Carmona, 2007) já que resulta na probabilidade do indivíduo pertencer a um determinado grupo, neste caso, o de usuários de uma ou mais drogas.

Na estimação e avaliação do modelo através da regressão logística, foi utilizado o método *stepwise backward* para selecionar as variáveis significativas para a predição.

O critério utilizado para a seleção foi o AIC, que considera a qualidade do modelo e a simplicidade do modelo. Em alguns casos, atribuiu-se pesos às observações para buscar melhores resultados quando a amostra possuía uma distribuição desbalanceada de usuários e não usuários. Outro ajuste testado, para buscar melhor acurácia e sensibilidade, foi a alteração do ponto de corte da probabilidade predita ao classificar os indivíduos como usuários ou não usuários.

Na implementação do método, foi utilizado o pacote **glm** do R, que executa uma gama de modelos lineares, com o parâmetro *family* como *binomial* identificando a regressão logística. Para cada droga ou grupo de droga, inicialmente, gerou-se um modelo utilizando todas as variáveis e obteve-se acurácia e sensibilidade. Em seguida, foi aplicada a função **stepAIC** do pacote **MASS** com o parâmetro *backward* para identificar o método para a seleção de variáveis. No caso das drogas ecstasy, álcool e o grupo de drogas estimulantes, foi adicionado o parâmetro *weight* para buscar melhores resultados em amostras desbalanceadas.

1.3.2 Árvore de Decisão

O modelo de Árvores de Decisão é um dos mais simples e intuitivos métodos de classificação através de Machine Learning. Focado na interpretabilidade do ajuste, as árvores de Decisão possuem esse nome devido ao fato da representação visual do seu processo decisório se assemelhar com a estrutura de uma Árvore. Partindo de um único ponto inicial, esse algoritmo busca classificar uma variável por meio de um conjunto de atributos que são utilizados para gerar uma sequência de ramificações, de tal forma que no final de cada ramo esteja a classe escolhida.

Apesar de ser um método com menor capacidade preditiva e com menos robustez do que vários dos demais algoritmos conhecidos de Machine Learning, as árvores de decisão possuem um papel importante quando realizamos análises focadas em inferência, pois permitem a criação de um modelo de fácil compreensão e interpretação. Além disso, as Árvores de decisão se destacam por serem um dos métodos que melhor lidam com a presença de atributos categóricos entre os dados. Por fim, elas possuem uma representação gráfica mesmo quando estamos trabalhando com dimensões maiores do que 3, uma proeza dificilmente alcançada pela grande maioria dos modelos.

Dessa forma, o uso desse método é facilmente justificável no presente trabalho: estamos em busca de um modelo que permita interpretações e possuímos um conjunto de dados inteiramente constituído por variáveis categóricas. Para aplicá-los, foi usado o pacote **rpart** do R, que implementa ideias do modelo conhecido como CART (Classification and Regression Tree), descrito em 1984 por Breiman et al. (2019). Nessa abordagem, a árvore começa a ser construída a partir da variável que melhor separa os dados em dois grupos - melhor separação, aqui, é definida como a ramificação que maximiza a redução da

impureza (medida de heterogeneidade) do nó segundo algum critério (em geral, índice de Gini ou índice de Informação). Após essa primeira separação, essas etapas são repetidas separadamente para cada subgrupo gerado. O processo segue de forma recursiva até não haver mais melhorias a serem feitas ou até que os subgrupos atinjam um tamanho mínimo. Nesse ponto, a Árvore final é gerada. A partir dela, ainda é possível aplicar um procedimento de validação cruzada para podá-la, reduzindo ainda mais o seu tamanho e a sua complexidade.

Assim, aplicou-se o procedimento acima para as drogas escolhidas utilizando-se a função **rpart** com o índice de Gini e tamanho mínimo dos subconjuntos como 20 observações. Obteve-se, assim, as matrizes de confusão e as medidas de acurácia, sensibilidade e especificidade do ajuste, além de uma figura representativa da árvore criada. Em todos os casos, as Árvores geradas foram posteriormente podadas utilizando- se a função **prune** e escolhendo-se o parâmetro de complexidade de tal forma que ele estivesse associado ao menor erro da validação cruzada que cumprisse a condição (erro relativo + desvio padrão da validação cruzada) < erro da validação cruzada. No caso de ecstasy e estimulantes, testou-se ainda o uso de uma matriz de perda que desse mais ênfase na classificação correta de usuários para corrigir o problema de amostras desbalanceadas.

Em todas as figuras associadas às árvores de Decisão, foram alterados os rótulo das variáveis pois os nomes originais eram muito longos. Para a variável Educação, adotamos os seguintes códigos: Left school = LS16, Some college or university, no certificate or degree = Uwodg, Professional certificate/diploma = Pct, University degree = Udg, Graduate degree = Gdg. Para a variável Gênero, utilizamos Male = M, Female = F. Para Idade, usamos 18-24 = J, 25-34 = A1, 35-44 = A2, 45+ = MI+. Para os Scores, utilizamos VeryLow = VL, Low = L, Average = A, High = H, VeryHigh = VH.

1.4 RESULTADOS

1.4.1 Cannabis

A distribuição de usuários e não usuários de maconha é bem balanceada, sendo 52% usuários, portanto, não foi necessário fazer um ajuste de pesos ou penalidades.

Regressão Logística

As variáveis selecionadas pelo método *stepwise backward* foram: **Idade (Age), Gênero (Gender), Educação (Education), Conscienciosidade (CScore), Abertura (OScore) e Neuroticismo (NScore)**. A Tabela 1.1 apresenta a matriz de confusão do modelo no conjunto de treino considerando apenas as variáveis selecionadas e utilizando o ponto de corte em 0,5. A acurácia do modelo foi de 76,46%, a sensibilidade 67,50% e a especificidade 86,47%. Os coeficientes estimados pelo modelo podem ser visualizados na Tabela 1.2.

Tabela 1.1: Matriz de confusão da regressão logística no conjunto de treino para a droga cannabis.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	569	102
Negativo	274	652

Tabela 1.2: Coeficientes da Regressão Logística para a Cannabis.

Variável	Coeficiente	Variável	Coeficiente
Intercept	0,3924	NScore: High	-0,7821
Age 25-34	-1,0330	NScore: Low	-0,3322
Age 35-44	-1,7719	NScore: Very Low	-0,4299
Age 45+	-2,2643	CScore: High	0,0837
Gender Male	0,9081	CScore: Low	0,6388
Education: LS16	1,3836	CScore: Very Low	1,0756
Education: Pct	0,7641	OScore: High	0,9083
Education: Uwodg	1,2540	OScore: Low	-1,0740
Education: Udg	0,4658	OScore: Very Low	-2,1160

Observa-se no gráfico da Figura 1.5 que, reduzindo o ponto de corte, é possível obter menos falsos negativos. Isso possibilita aumentar a sensibilidade, que é a métrica de interesse do estudo, pois indica a probabilidade do algoritmo classificar um usuário corretamente.

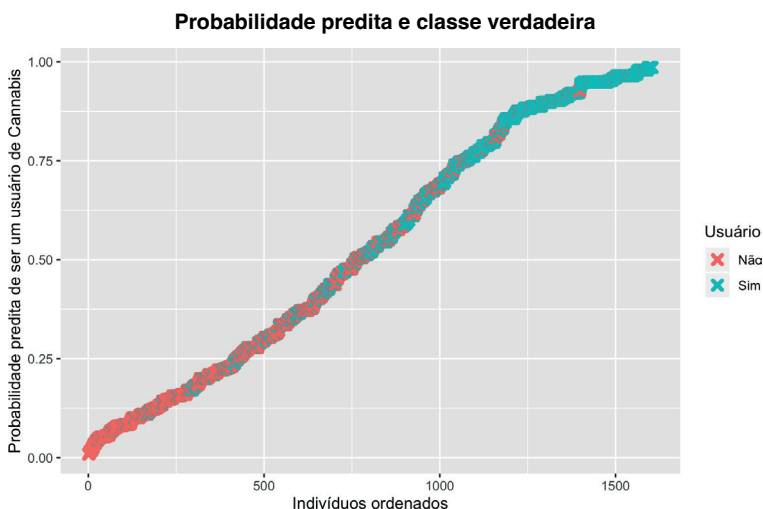



Figura 1.5: Gráfico da probabilidade predita e respectiva classe das observações dos usuários de maconha.

Optando por um ponto de corte de 0,3, tem-se que, ao classificar como usuários indivíduos com probabilidade predita acima de 0,3, há um aumento da sensibilidade para 71,41%, aumentando a acurácia para 77,27%, mas com redução na especificidade (83,82%), como pode ser observado na Tabela 1.3. Como o interesse é avaliar as duas primeiras medidas, considera-se que a classificação utilizando este ponto de corte é mais adequada.

Tabela 1.3: Matriz de confusão da regressão logística no conjunto de treino para a droga cannabis com ponto de corte em 0,3.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	602	122
Negativo	241	632

Aplicando os ajustes escolhidos no conjunto de teste, observa-se que os valores de acurácia, sensibilidade e especificidade ficaram próximos. A sensibilidade aumentou para 75%, a acurácia se manteve muito próxima (76,43)% e especificidade reduziu para 78,03%, como observado nas Tabelas 1.4 e 1.5.

Tabela 1.4: Matriz de confusão da regressão logística no conjunto de teste para a droga cannabis com ponto de corte em 0,3.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	111	29
Negativo	37	103

Tabela 1.5: Métricas de performance da regressão logística para a droga cannabis.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino (corte 0,5)	76,46%	67,50%	86,47%
Treino (corte 0,3)	77,27%	71,41%	83,82%
Teste (corte 0,3)	76,43%	75,00%	78,03%

Avaliando os valores de *odds ratio* das variáveis, apresentados na Tabela 1.6, há a indicação de que pessoas com idade acima de 45 anos possuem 90% menos chance de serem usuárias quando comparadas com pessoas entre 18 e 24 anos. Ainda, percebe-se que indivíduos com alto *score* na característica de personalidade Abertura (O) possuem 2,48 vezes mais chance de serem usuários do que os com *score* mediano e os com *score* muito baixo na característica Conscienciosidade (C) 2,93 vezes mais. É interessante ressaltar também que a chance é 3,99 vezes maior de uma pessoa ser usúaria de maconha ao parar de estudar aos 18 anos ou menos do que quem possui um diploma de pós graduação.

Tabela 1.6: Odds Ratio da Regressão Logística para a Cannabis.

Variável	Coeficiente	Variável		Coeficiente
Age 25-34	0,36	NScore:	Low	0,72
Age 35-44	0,17	NScore:	Very Low	0,65
Age 45+	0,10	CScore:	High	1,09
Gender Male	2,48	CScore:	Low	1,89
Education: LS16	3,99	CScore:	Very Low	2,93
Education: Pct	2,15	OScore:	High	2,48
Education: Uwodg	3,50	OScore:	Low	0,34
Education: Udg	1,59	OScore:	Very Low	0,12
NScore: High	0,46			

Árvore de Decisão

Aplicando-se a função **rpart**, obteve-se um modelo com as seguintes variáveis: **Idade (Age), Conscienciosidade (CScore), Gênero (Gender), Abertura (OScore)**. O modelo foi capaz de obter acurácia, sensibilidade e especificidade consideravelmente altas tanto para o conjunto de treino quanto o de teste, com todas as métricas ficando acima dos 70%. Esses resultados podem ser observados em maiores detalhes nas Tabelas 1.7 e 1.8. Uma visualização do processo decisório da Árvore produzida pode ser observada na Figura 1.6.

Tabela 1.7: Métricas de performance da árvore de Decisão para a droga cannabis.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino	75,39%	74,02%	76,92%
Teste	71,43%	71,62%	71,21%

Tabela 1.8: Tabela de confusão da árvore de Decisão para a droga cannabis.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	106	38
Negativo	42	94

Árvore de Decisão - Cannabis

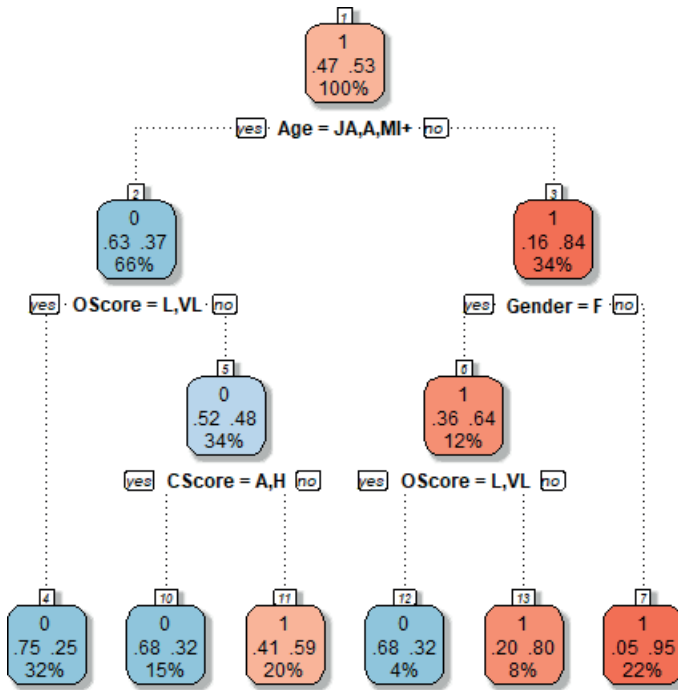



Figura 1.6: Visualização da árvore de Decisão para a droga cannabis.

Realizados esses passos, aplicou-se o método de poda da Árvore. A Árvore podada acabou resultando na própria árvore original, indicando que o modelo anterior já é o mais parcimonioso que conseguimos obter para essa análise.

1.4.2 Ecstasy

Na amostra coletada, tem-se que apenas 27,42% dos indivíduos são usuários de Ecstasy. Este desbalanceamento tem efeito na sensibilidade do modelo, portanto, buscou-se corrigi-lo com técnicas de atribuição de pesos e penalidades.

Regressão Logística

As variáveis selecionadas foram: **Idade (Age), Gênero (Gender), Educação (Edu- cation) e Abertura (OScore)**. Observa-se na Tabela 1.9 a matriz de confusão do modelo no conjunto de treino apenas com as variáveis selecionadas e o ponto de corte em 0,5. A acurácia do modelo foi relativamente alta, de 74,77%, e a especificidade (93,71%) também. Porém, a sensibilidade foi apenas 24,49%, pois o banco de dados possui muito menos usuários do que não usuários e o modelo estimado acaba atribuindo baixas probabilidades aos usuários.

Tabela 1.9: Matriz de confusão da regressão logística no conjunto de treino para a droga ecstasy com ponto de corte em 0,5.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	107	73
Negativo	330	1.087

Para buscar corrigir o desbalanceamento, executou-se a regressão logística com seleção de variáveis no conjunto de treino, mas adicionando peso **3** para usuários e **1** para não usuários. As variáveis selecionadas aumentaram, sendo incluídas **Conscienciosidade (CScore) e Socialização (AScore)**. Observa-se no gráfico da Figura 1.7 que o ponto de corte em torno de 0,5 é adequado e optou-se por mantê-lo.

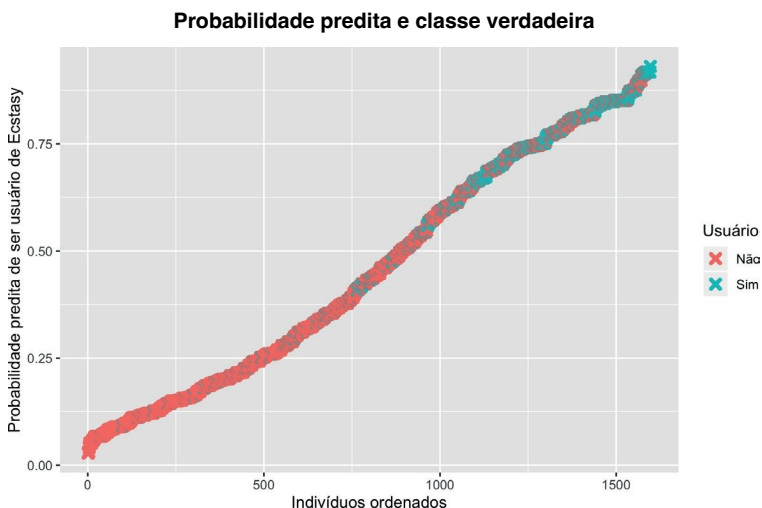



Figura 1.7: Gráfico da probabilidade predita e respectiva classe das observações dos usuários de ecstasy.

As Tabelas 1.10 e 1.11 apresentam, respectivamente, a matriz de confusão do modelo no conjunto de treino e os coeficientes estimados para esse modelo. Observamos na Tabela 1.10 que a especificidade reduziu para 78,88%. Porém, a acurácia e a sensibilidade, que são as métricas de interesse, aumentaram para, respectivamente, 76,08% e 68,65%. Utilizando modelo ajustado com pesos e ponto de corte em 0,5 no conjunto de teste, pode-se observar que os valores de acurácia, sensibilidade e especificidade ficaram próximos, tendo a sensibilidade reduzido para 65,79% e a acurácia e especificidade aumentado para 77,14% e 81,37%, respectivamente, como observado nas Tabelas 1.12 e 1.13.

Tabela 1.10: Matriz de confusão do conjunto de treino da regressão logística com pesos para a droga ecstasy com ponto de corte em 0,5.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	300	245
Negativo	137	915

Tabela 1.11: Coeficientes da Regressão Logística para o Ecstasy.

Variável	Coeficiente	Variável		Coeficiente
Intercept	0,0742	CScore:	High	-0,2756
Age 25-34	-0,6736	CScore:	Low	0,2892
Age 35-44	-1,7943	CScore:	Very Low	0,2665
Age 45+	-2,3933	AScore:	High	-0,5383
Gender Male	0,6453	AScore:	Low	0,2423
Education: LS16	0,4383	AScore:	Very Low	0,4482
Education: Pct	0,3875	OScore:	High	0,6693
Education: Uwodg	0,4939	OScore:	Low	-0,6911
Education: Udg	0,0995	OScore:	Very Low	-0,5869

Tabela 1.12: Matriz de confusão do conjunto de teste da regressão logística com pesos para a droga ecstasy com ponto de corte em 0,5.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	50	38
Negativo	26	166

Tabela 1.13: Métricas de performance da regressão logística para a droga ecstasy.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino (corte 0,5)	74,77%	24,49%	93,71%
Treino com pesos (corte 0,5)	76,08%	68,65%	78,88%
Teste com pesos (corte 0,5)	77,14%	65,79%	81,37%

Através dos valores de *odds ratio* das variáveis, apresentados na Tabela 1.14, sugere-se que o *score* alto na característica Abertura (O) aumenta a chance de uma pessoa ser usúaria de Ecstasy em 1,95 vezes, o *score* muito baixo em Socialização (A) aumenta em 1,57 vezes, e o *score* muito baixo em Conscienciosidade (C) aumenta em 1,31 vezes, quando comparados aos indivíduos com *scores* medianos. Ainda, homens têm 1,91 vezes mais chance de serem usuários do que mulheres.

Tabela 1.14: Odds Ratio Regressão Logística para o Ecstasy.

Variável	Coeficiente	Variável		Coeficiente
Age 25-34	0,51	CScore:	Low	1,34
Age 35-44	0,17	CScore:	Very Low	1,31
Age 45+	0,09	AScore:	High	0,58
Gender Male	1,91	AScore:	High	0,58
Education: LS16	1,55	AScore:	Low	1,27
Education: Pct	1,47	AScore:	Very Low	1,57
Education: Uwodg	1,64	OScore:	High	1,95
Education: Udg	1,10	OScore:	Low	0,50
CScore: High	0,76	OScore:	Very Low	0,56

Árvores de Decisão

Para a droga ecstasy, obteve-se um modelo com as seguintes variáveis: **Idade (Age), Conscienciosidade (CScore), Gênero (Gender), Abertura (OScore), Extor-versão (Escore)**. Apesar da acurácia do ajuste ser relativamente alta, em cerca de 77%, a sensibilidade do ajuste fica consideravelmente baixa, na faixa dos 40%. Os resultados compilados estão na Tabela 1.15 e a matriz de confusão para o conjunto de teste pode ser observada em 1.16. A visualização da árvore produzida pode ser observada na Figura 1.8.

Tabela 1.15: Métricas de performance da árvore de Decisão para a droga ecstasy.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino	76,58%	43,71%	88,97%
Teste	77,14%	40,79%	90,69%

Devido à baixa sensibilidade obtida no ajuste, decidiu-se refazer a análise utilizando uma matriz de perda que focasse mais na predição correta dos usuários - ou seja, o erro de classificar usuários como não usuários foi considerado duas vezes pior que cometer o erro inverso. Dessa vez, o modelo obtido continha as variáveis **Idade (Age), Gênero (Gender), Abertura (OScore)**. A acurácia do ajuste baixou apenas 3%, permanecendo acima dos 70%, enquanto a sensibilidade subiu em quase 30% - e esse resultado é exatamente o que se desejava obter aplicando a penalização sobre os erros. A Tabela 1.17 apresenta os resultados obtidos, enquanto a matriz de confusão está na Tabela 1.18. A visualização da nova árvore pode ser vista na Figura 1.9.

Tabela 1.16: Tabela de confusão do conjunto de teste da árvore de Decisão para a droga ecstasy.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	31	19
Negativo	45	185

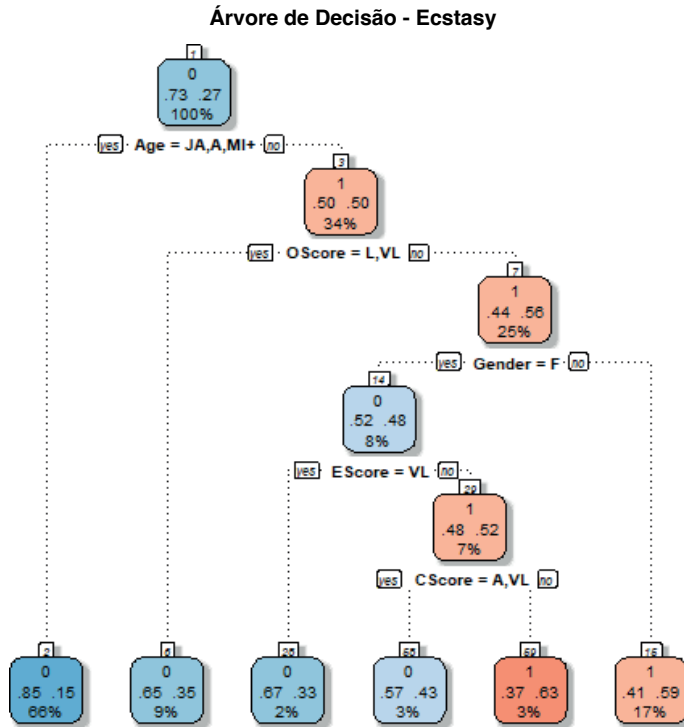



Figura 1.8: Visualização da árvore de Decisão para a droga ecstasy.

Tabela 1.17: Métricas de performance da árvore de Decisão com função perda modificada para a droga ecstasy.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino	73,32%	76,20%	72,24%
Teste	74,64%	69,34%	76,47%

Tabela 1.18: Tabela de confusão do conjunto de teste da árvore de Decisão com função perda modificada para a droga ecstasy.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	53	48
Negativo	23	156

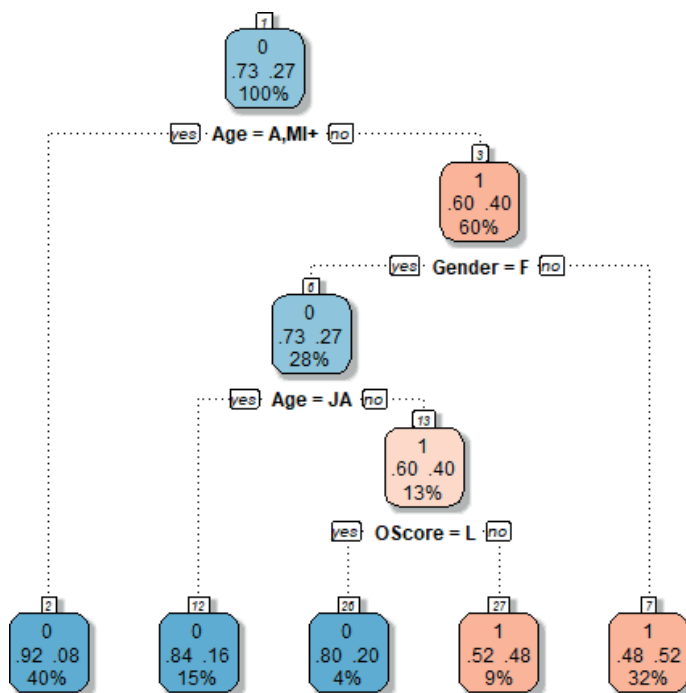



Figura 1.9: Visualização da árvore de Decisão com função perda modificada para a droga ecstasy.

Por fim, aplicou-se o método de poda sobre a Árvore com função perda modificada. A árvore podada obtida corresponde a um modelo apenas com as variáveis **Idade (Age), Gênero (Gender)**. Apesar da especificidade da árvore podada aumentar cerca de 5%, a acurácia aumenta menos de 1% e a sensibilidade do ajuste baixa em torno de 15%, como pode ser visto na Tabela 1.19. Como as últimas métricas são consideradas mais essenciais para a análise, optou-se por não adotar a Árvore podada e permanecer apenas com a árvore penalizada.

Tabela 1.19: Métricas de performance da árvore de Decisão para a droga ecstasy.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino	73,76%	60,41%	78,79%
Teste	75,00%	56,58%	81,86%

1.4.3 Álcool

Foi observado que a amostra utilizada para avaliar se um indivíduo é ou não usúario de álcool é exageradamente desigual: apenas 7% dos indivíduos da amostra não são usuários. Isso torna a predição do modelo muito difícil, de tal forma que nenhum dos modelos estimados possui boas propriedades.

Regressão Logística

Na regressão logística, mesmo atribuindo um peso oito vezes maior aos não usuários, não é possível ter uma acurácia aceitável, de forma que classificar sempre o indivíduo como usuário tem melhores resultados do que executar o modelo, como mostra a Tabela 1.20. Os coeficientes estimados pelo modelo são apresentados na Tabela 1.21.

Tabela 1.20: Métricas de performance da regressão logística para a droga álcool.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino	61,09%	62,60%	60,97%
Teste	56,58%	75,00%	55,17%

Tabela 1.21: Coeficientes da Regressão Logística para o Álcool.

Variável	Coeficiente	Variável		Coeficiente
Intercept	0,9699	NScore:	Very Low	0,3909
Age 25-34	-0,2805	CScore:	High	0,6059
Age 35-44	-0,6758	CScore:	Low	0,3635
Age 45+	-0,9388	CScore:	Very Low	0,5160
Gender Male	-0,2370	AScore:	High	-0,4261
Education: LS16	-0,6666	AScore:	Low	0,1961
Education: Pct	-0,5805	AScore:	Very Low	0,2040
Education: Uwodg	-0,3705	EScore:	High	13,5985
Education: Udg	-0,2764	EScore:	Low	-0,2406
NScore: High	0,2320	EScore:	Very Low	-0,2219
NScore: Low	0,2912			

Árvore de Decisão

Para o álcool, houveram dificuldades de ajustar uma árvore de Decisão: devido a altíssima proporção de usuários de álcool na base, não foi possível gerar modelos. O algoritmo não consegue obter nenhuma partição que diminua a impureza, então ele retorna apenas uma raiz. A única maneira de se gerar um modelo de Árvore de decisão nesse caso é forçando a criação de uma árvore sobreajustada que utiliza todos os atributos disponíveis. A árvore gerada desta forma é mostrada na Figura 1.10, onde nota-se com clareza que o modelo gerado é de péssima qualidade, sendo inútil para utilização no problema de classificação. As métricas do ajuste forçado são apresentados na Tabela 1.22. Observe que as métricas são incrivelmente altas, mas isso deve-se apenas ao viés inerente da base.

Árvore de Decisão - Álcool

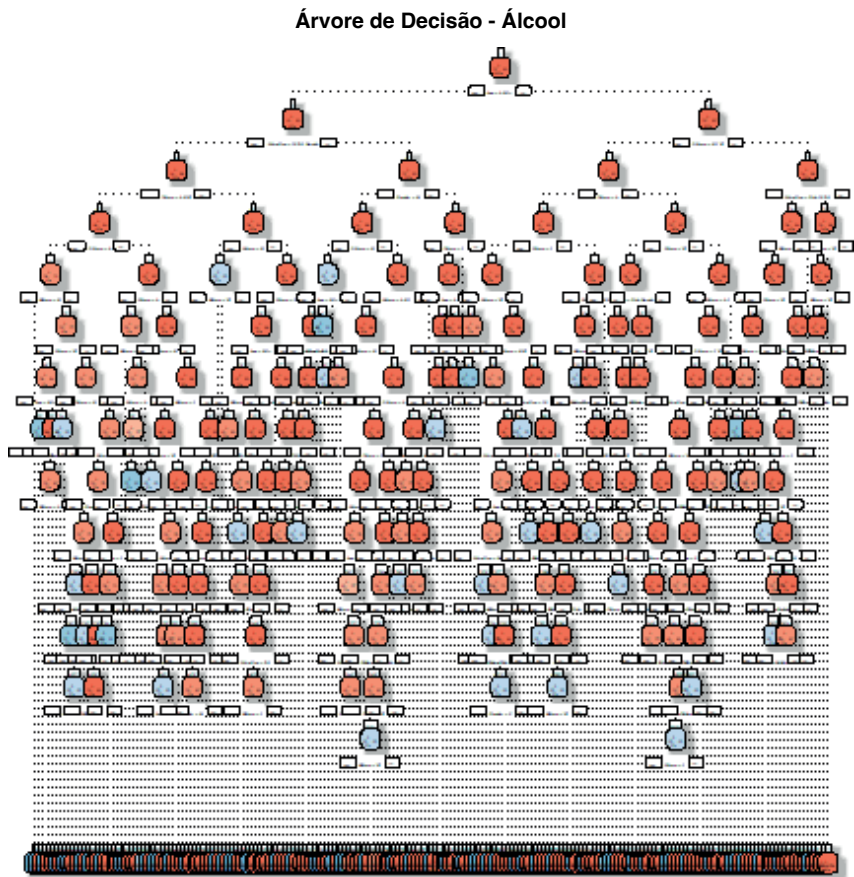



Figura 1.10: Visualização da árvore de Decisão para a droga álcool.

Tabela 1.22: Métricas de performance da árvore de Decisão para a droga álcool.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino	97,30%	99,80%	65,22%
Teste	85,76%	91,19%	15,00%

1.4.4 Grupo de Drogas Estimulantes

A distribuição de usuários e não usuários de drogas estimulantes é de 67,5% usuários e 32,5% não usuários, o que nos leva a utilizar pesos e penalidades para buscar melhores resultados.

Regressão Logística

As variáveis selecionadas foram: **Idade (Age), Gênero (Gender), Educação (Education), Abertura (OScore), Conscienciosidade (CScore) e Socialização (AScore)**. Ajustando o modelo no conjunto de treino com as variáveis selecionadas e o ponto de corte em 0,5, obteve-se 70,93% de acurácia, 93,22% de especificidade, mas apenas 24,51% de sensibilidade, como observado nas Tabelas 1.23 e 1.24.

Tabela 1.23: Matriz de confusão da regressão logística no conjunto de treinamento para o grupo de drogas estimulantes.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	127	73
Negativo	391	1.005

Tabela 1.24: Métricas de performance da regressão logística para a drogas estimulantes.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino (corte 0,5)	70,93%	24,51%	93,22%
Treino com pesos (corte 0,5)	72,81%	52,90%	82,37%
Treino com pesos (corte 0,35)	72,74%	59,27%	79,22%
Teste com pesos (corte 0,35)	70,82%	63,74%	74,21%

Similarmente ao método empregado para a droga ecstasy, atribuiu-se um peso maior aos usuários de drogas estimulantes para buscar melhores resultados na métrica sensibilidade. Para seguir a proporção da amostra, utilizou-se peso **2** para usuários e **1** para não usuários. As variáveis selecionadas se mantiveram as mesmas e foi possível aumentar a acurácia e a sensibilidade para 72,81% e 52,90%, respectivamente, como apresentado nas Tabelas 1.25 e 1.24.

Tabela 1.25: Tabela de confusão da regressão logística no conjunto de treino com pesos e ponto de corte em 0,5 para o grupo de drogas estimulantes.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	274	190
Negativo	244	888

Ainda, o gráfico apresentado na Figura 1.11 sugere que é possível aumentar a métrica de interesse reduzindo o ponto de corte.

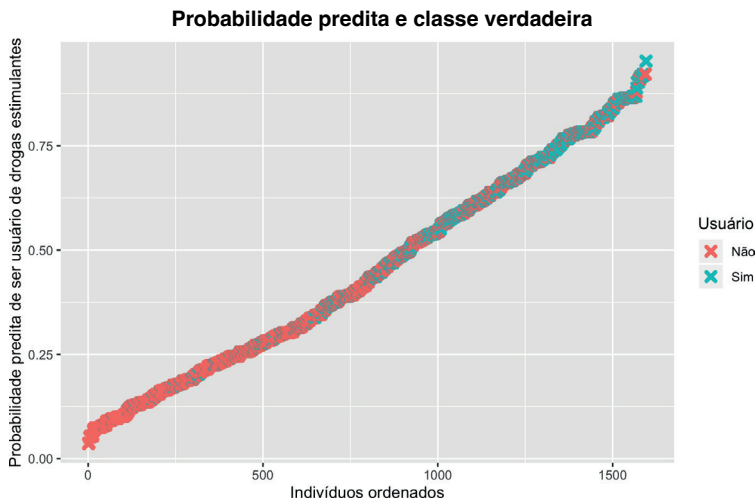



Figura 1.11: Gráfico da probabilidade predita e respectiva classe das observações dos usuários de drogas estimulantes.

Classificando como usuários os indivíduos com probabilidade predita maior ou igual a 0,35, obteve-se um aumento na sensibilidade para 59,27%, sem comprometer a acurácia, que ficou muito próxima (72,74%) e comprometendo a especificidade, que reduziu para 79,22%, como apresentado nas Tabelas 1.26 e 1.24. Os coeficientes estimados pelo modelo podem ser visualizados na Tabela 1.27.

Tabela 1.26: Tabela de confusão da regressão logística com pesos no conjunto de treino e ponto de corte em 0,35 para o grupo de drogas estimulantes.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	307	224
Negativo	211	854

Tabela 1.27: Coeficientes da Regressão Logística para o grupo de Drogas Estimulantes.

Variável	Coeficiente			
Intercept	-0,6014	CScore:	High	0,3350
Age 25-34	-0,2366	CScore:	Low	0,3863
Age 35-44	-1,0288	CScore:	Very Low	0,9580
Age 45+	-1,6973	AScore:	High	-0,3841
Gender Male	0,5969	AScore:	Low	0,3295
Education: LS16	0,4797	AScore:	Very Low	0,9352
Education: Pct	0,3919	OScore:	High	0,5459
Education: Uwodg	0,5741	OScore:	Low	-0,6141
Education: Udg	0,0581	OScore:	Very Low	-0,6744

A Tabela 1.28 apresenta a matriz de confusão com os resultados da aplicação do modelo estimado no conjunto de teste, com peso 2 atribuído a usuários de drogas estimulantes e 1 para não usuários e ponto de corte em 0,35. Observa-se que acurácia ficou em 70,82%, a sensibilidade aumentou para 63,74% e a especificidade reduziu para 74,21%. Considerou-se um bom ajuste, dado que as métricas de interesse foram satisfatórias.

Tabela 1.28: Tabela de confusão da regressão logística com pesos no conjunto de teste e ponto de corte em 0,35 para o grupo de drogas estimulantes.

Teste	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	58	49
Negativo	33	141

Percebe-se, através da avaliação dos valores de *odds ratio* das variáveis, apresentados na Tabela 1.29, que indivíduos com *scores* de Conscienciosidade (C) e Socialização (A) muito baixos possuem 2,61 e 2,55 vezes, respectivamente, mais chance de serem usuários de drogas estimulantes quando comparados a indivíduos com *score* mediano. Ainda, pessoas com mais de 45 anos de idade têm 82% menos chance de serem usuários do que os jovens entre 18 e 24 anos, e homens têm 1,82 vezes mais possibilidade de serem usuários do que mulheres.

Tabela 1.29: Odds Ratio da Regressão Logística para o grupo de Drogas Estimulantes.

Variável	Coeficiente	Variável	Coeficiente
Age 25-34	0,79	CScore: High	1,40
Age 35-44	0,36	CScore: Low	1,47
Age 45+	0,18	CScore: Very Low	2,61
Gender Male	1,82	AScore: High	0,68
Education: LS16	1,62	AScore: Low	1,39
Education: Pct	1,48	AScore: Very Low	2,55
Education: Uwodg	1,78	OScore: High	1,73
Education: Udg	1,06	OScore: Low	0,54
		OScore: Very Low	0,51

Árvore de Decisão

Para o conjunto de drogas estimulantes, obteve-se um modelo com as variáveis **Idade (Age), Socialização (AScore), Educação (Education), Gênero (Gender), Neuroticismo (NScore), Abertura (OScore)**. As matrizes de confusao para os conjuntos de treinamento e teste podem ser observadas nas Tabelas 1.30 e 1.31, respectivamente. Assim como no caso do ecstasy, a acurácia foi consideravelmente alta, na casa dos 70%, mas a sensibilidade do ajuste não ultrapassa os 40%, tal como pode ser visto na Tabela 1.30. A árvore produzida pode ser visualizada na Figura 1.12.

Tabela 1.30: Métricas de performance da árvore de Decisão para o grupo de drogas estimulantes.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino	74,87%	40,15%	91,56%
Teste	71,53%	35,16%	88,95%

Tabela 1.31: Tabela de confusão do conjunto de teste da árvore de Decisão para o grupo de drogas estimulantes.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	32	21
Negativo	59	169

Seguindo as mesmas etapas realizadas para o ecstasy, a análise foi refeita considerando o erro de classificar usuários como não usuários como sendo duas vezes pior que cometer o erro inverso. Dessa vez, o modelo obtido continha as variáveis **Idade (Age), Socialização (AScore), Gênero (Gender), Neuroticismo (NScore), Abertura (OScore)**. Apesar da acurácia do ajuste cair quase 5% e da especificidade dele cair mais de 20%, a sensibilidade, que é nosso foco, sobe mais de 35%, como mostrado na Tabela 1.32. A matriz de confusão pode ser observada na Tabela 1.33, enquanto a visualização da nova árvore pode ser vista na Figura 1.13.

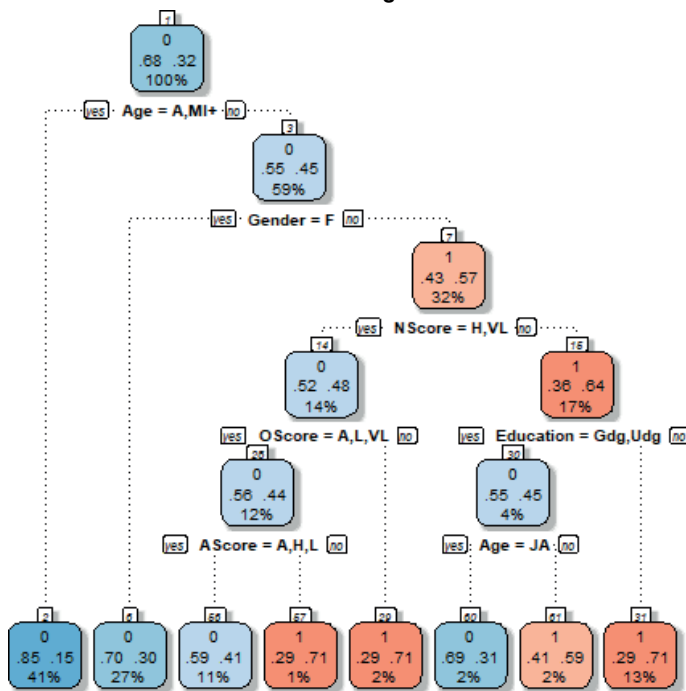



Figura 1.12: Visualização da Árvore de Decisão para o grupo de drogas estimulantes

Tabela 1.32: Métricas de performance da árvore de Decisão com função perda modificada para o grupo de drogas estimulantes.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino	71,05%	74,90%	69,20%
Teste	66,90%	76,92%	62,10%

Por fim, aplicou-se o método de poda sobre a Árvore com função perda modificada. A árvore podada obtida corresponde a um modelo apenas com as variáveis **Idade (Age), Gênero (Gender), Abertura (OScore)**. Analisando-se as Tabelas 1.34 e 1.35, nota-se que, apesar de termos retirado duas variáveis do modelo, nenhuma das métricas de performance caiu consideravelmente, indicando que aqui o uso da árvore podada pode ser adequado. A árvore final pode ser observada na Figura 1.14.

Tabela 1.33: Tabela de confusão do conjunto de teste da árvore de Decisão com função perda modificada para o grupo de drogas estimulantes.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	70	72
Negativo	21	118

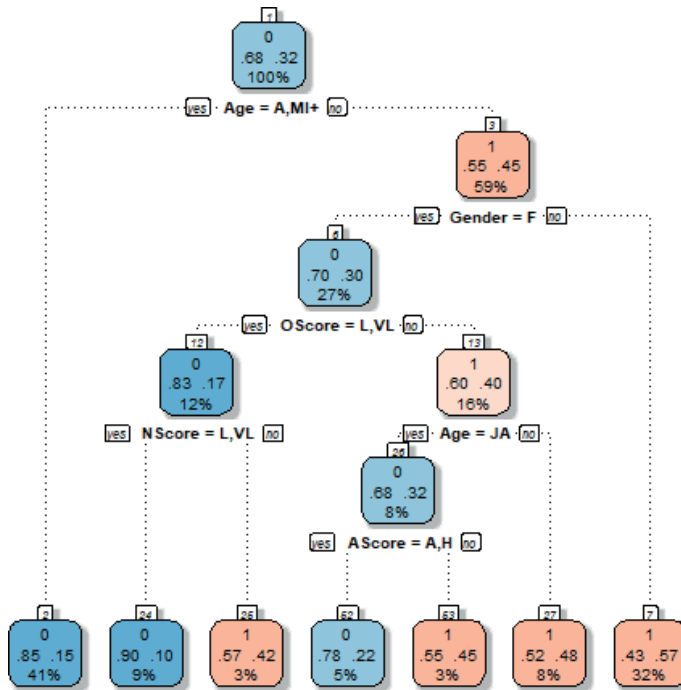



Figura 1.13: Visualização da Árvore de decisão com função perda modificada para o grupo de drogas estimulantes.

Tabela 1.34: Métricas de performance da árvore de Decisão para o grupo de drogas estimulantes.

Conjunto	Acurácia	Sensibilidade	Especificidade
Treino	68,86%	74,71%	66,05%
Teste	64,77%	74,72%	60,00%

Tabela 1.35: Tabela de confusão do conjunto de teste da árvore de Decisão podada para o grupo de drogas estimulantes.

	Valor Verdadeiro	
Valor Predito	Positivo	Negativo
Positivo	68	76
Negativo	23	114

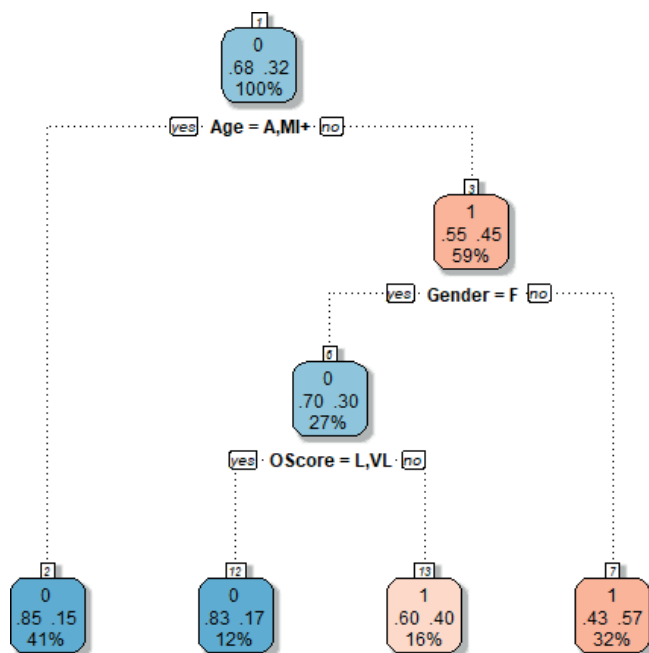



Figura 1.14: Visualização da árvore de Decisão podada para o grupo de drogas estimulantes.

1.4.5 Comparativo dos Resultados

Finalizada a aplicação dos métodos de regressão logística e de árvores de Decisão para todas as drogas escolhidas, podemos analisar os resultados para comparar a performance dos métodos e procurar por padroes nos resultados. Um compilado dos resultados finais obtidos pode ser observado na Tabela 1.36. Os resultados para o álcool foram desprezados devido à péssima qualidade dos ajustes.

Tabela 1.36: Acurácia, sensibilidade, especificidade e número de variáveis (k) do modelo obtido para os métodos de regressão logística e árvore de Decisão.

	Regressão Logística				**Árvore de Decisão**			
	Acur.	Sens.	Especif.	k	Acur.	Sens.	Especif.	k
Cannabis								
Treino	77,27%	71,41%	83,82%	6	75,39%	74,02%	76,92%	4
Teste	76,43%	75,00%	78,03%	6	71,43%	71,62%	71,21%	4
Ecstasy								
Treino	76,08%	68,65%	78,88%	6	73,32%	76,20%	72,24%	3
Teste	77,14%	65,79%	81,37%	6	74,64%	69,34%	76,47%	3
Estimul.								
Treino	72,74%	59,29%	79,22%	6	68,86%	74,71%	66,05%	3
Teste	70,82%	63,74%	74,21%	6	64,77%	74,72%	60,00%	3

A partir dos resultados da Tabela 1.36, é possível notar que os métodos aplicados possuem as suas particularidades. A regressão logística performa consistentemente melhor em termos de acurácia e de especificidade, enquanto a árvore de Decisão em geral alcança resultados melhores para a sensibilidade e consegue gerar modelos mais parcimoniosos. Apesar disso, os resultados dos dois modelos ficaram consideravelmente próximos, sendo ambos válidos para serem aplicados nessa análise.

Além dessas questões, uma observação mais aprofundada dos problemas de classificação estudados nos permite avaliar a aplicabilidade desses modelos sobre distribuições desbalanceadas. As análises mais simples de serem realizadas foram as da droga cannabis, cuja distribuição era bem balanceada. Nos casos em que a distribuição das classes fica na faixa dos 70% - 30%, o que corresponde ao ecstasy e ao conjunto de drogas estimulantes, as modelagens precisaram levar em conta o uso de técnicas extras (uso de pesos/custos mais elevados para erros de classificação) para alcançar uma performance adequada. Por fim, no caso do álcool, em que temos uma amostra extremamente desbalanceada (com uma das classes tendo uma predominância de mais de 90%), nenhum dos modelos se mostrou efetivo, evidenciando que existem situações onde os problemas pré-existentes no banco de dados não conseguem ser superados pelos algoritmos.

1.4.6 Resultados por Atributo

Além de analisarmos os resultados do ponto de vista dos métodos utilizados, podemos analisá-los da perspectiva dos atributos utilizados. Em outras palavras, podemos avaliar quais variáveis estavam presentes em cada modelo e quais dos seus níveis estão ligados ao uso de determinada droga. Um compilado da presença de cada uma das variáveis nos diferentes ajustes finais obtidos pode ser encontrado na Tabela 1.37. Devido a péssima qualidade do ajuste para o álcool, iremos desconsiderá-lo nessa análise e focar apenas nas drogas ilícitas.

Tabela 1.37: Presença dos atributos na seleção de variáveis dos métodos de regressão logística (RL) e árvore de Decisão (AD).

	Idade	Educação	Gênero	Score				
				N	E	O	A	C
Cannabis								
RL	X	X	X	X		X		X
AD	X		X			X		X
Ecstasy								
RL	X	X	X			X	X	X
AD	X		X			X		
Estimulantes								
RL	X	X	X			X	X	X
AD	X		X			X		

Analisando cada uma das variáveis de maneira isolada e levando em consideração todos os modelos gerados, podemos ter um panorama do perfil de usuário de drogas. Começando pelos indicadores sociais, temos que Idade costuma ser um fator relevante pois os mais jovens (entre 18 e 24 anos) possuem uma tendência mais elevada de serem usuários de drogas em comparação com os indivíduos de idade mais avançada, em especial considerando a faixa acima dos 45 anos. Da mesma forma, os homens aparecem como um grupo de maior risco, chegando a ter quase o dobro de chance de serem usuários de drogas do que as mulheres para algumas das drogas. É importante ressaltar, contudo, que jovens na faixa dos 18 aos 24 anos são maioria na amostra, o que pode interferir no resultado.

Educação já é um atributo de comportamento mais instável: ele aparece em todas as análises de regressão logística, mas em nenhum dos modelos finais das árvores de decisão. Nos casos em que ela aparece no modelo, a tendência indica que pessoas com um nível mais elevado de educação possuem uma chance significativamente menor de serem usuários de droga, em especial em comparação com as pessoas que abandonaram a escola ou que não possuem nenhum diploma.

Em relação aos Scores do NEO-FFI-R, tem-se que os principais fatores são o OScore (Abertura) e o CScore (Conscienciosidade). Para o Score que mede a Abertura, temos que o uso de drogas está bastante relacionado com o fato de possuir um valor mais elevado (médio ou alto) nessa métrica. No caso da Conscienciosidade, o inverso ocorre: o uso de drogas está ligado com valores baixos de Score nessa métrica.

Os demais Scores já possuem uma presença mais esporádica: EScore, a medida associada à Extroversão, nem sequer aparece nos modelos finais, sendo portanto uma variável irrelevante para determinar o uso de drogas. Já o Score de Socialização (AScore), quando aparece, indica que indivíduos com pontuações baixas nessa categoria possuem uma maior tendência a serem usuários de drogas. Por fim, o score de Neuroticismo (NScore) aparece em apenas um dos modelos e possui um comportamento menos definido do que os demais: a classe com maior chance de uso de drogas é a com pontuações de valores médios. Essa instabilidade é justificada em boa parte pela própria distribuição dos dados: observando novamente a Figura 1.3, nota-se que esse Score é o que mais foge da distribuição normal, estando extremamente concentrado em torno dos valores mais baixos.

1.5 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo observar características associadas ao uso de drogas específicas, utilizado para isto as técnicas de classificação de regressão logística e de Árvores de decisão. Em uma análise mais detalhada dos métodos, foi possível notar que os modelos ajustados apresentaram acurácia e especificidade levemente superiores na modelagem por regressão e acima de 70% para as drogas cannabis e ecstasy, mantendo a performance quando comparados treino e teste. Para as drogas estimulantes, a árvore de decisão foi que apresentou pior desempenho com acurácia e especificidade respectivamente de 68,68% e 66,05% no treino e 64,77% e 60,05% no teste. Em relação a sensibilidade, as árvores de decisão apresentaram resultados ligeiramente superiores a regressão logística. Pode-se concluir que os modelos obtidos, tanto na regressão logística, quanto nas árvores de decisão, são satisfatórios na classificação de usuários para as drogas cannabis, ecstasy e para o grupo de drogas estimulantes. Para o álcool não foi possível estimar um modelo adequado, dada a proporção da amostra obtida.

A indicação final é a utilização dos modelos que classifiquem melhor em relação a sensibilidade de forma a não prejudicar a acurácia, pois estamos realizando análises com objetivo inferencial e com foco na modelagem dos usuários de drogas. Dessa forma, os melhores modelos seriam o modelo de regressão para cannabis, enquanto ecstasy e o grupo de drogas estimulantes são ajustados de maneira mais adequada pela árvore de decisão. Contudo, é importante ressaltar que os modelos tiveram performances consideravelmente similares, e que escolher um deles em detrimento do outro não acarreta em grandes perdas de performance.

Em relação ao perfil dos usuários de drogas, as variáveis idade, gênero e Oscore se mostraram como as mais importantes, entrando em todos os modelos propostos. Considerando os diferentes modelos, o perfil geral do usuário de drogas que se observou caracteriza esses indivíduos como sendo jovens, predominantemente homens e com nível educacional mais baixo. Quanto aos traços de personalidade do NEO-FFI-R, os resultados indicam que o consumo de drogas está ligado com baixos escores de Conscienciosidade (C) e de Socialização (A) e com altos escores de Abertura(O). Esses achados estão de acordo com a literatura sobre o assunto Fehrman et al. (2015b).

Por fim, é importante ressaltar que os dados coletados podem estar viesados, dado a estratégia adotada de amostragem não probabilística, impossibilitando a generalização dos resultados para a população. Ainda assim, os resultados obtidos são interessantes tanto para a geração de *insights* para possíveis ações de prevenção, quanto para próximas pesquisas.

Sugere-se como análises futuras a exploração de tais problemas de classificação para as demais drogas que não foram abordadas nesse estudo, de tal forma a termos um espectro mais completo do perfil de usuários de diferentes drogas. Além disso, a repetição dessas mesmas análises em amostra não viesadas dos dados é desejável, de tal forma a nos permitir determinar com mais clareza o quanto cada variável de fato influencia no uso de drogas.

Apêndice: tabelas

Tabela 1.38: Frequências das variáveis sexo, idade, educação, país e grupo étnico nas amostras de interesse.

		Cannabis		**Ecstasy**		**Álcool**		**Estimulantes**	
	Amostra	Usuário	Não Usuário	Usuário	Não Usuário	Usuário	Não Usuário	Usuário	Não Usuário
	n = 1.877	991 (53%)	886 (47%)	513 (27%)	1.364 (73%)	1.742 (93%)	135 (7%)	609 (32%)	1.268 (68%)
Sexo									
Feminino	937	359 (38%)	578 (62%)	166 (18%)	771 (82%)	873 (93%)	64 (7%)	210 (22%)	727 (78%)
Masculino	940	632 (67%)	308 (33%)	347 (37%)	593 (63%)	869 (92%)	71 (8%)	399 (42%)	541 (58%)
Idade									
18-24	637	530 (83%)	107 (17%)	321 (50%)	316 (50%)	611 (96%)	26 (4%)	329 (52%)	308 (48%)
25-34	480	237 (49%)	243 (51%)	126 (26%)	354 (74%)	453 (94%)	27 (6%)	167 (35%)	313 (65%)
35-44	355	122 (34%)	233 (66%)	44 (12%)	311 (88%)	321 (90%)	34 (10%)	66 (19%)	289 (81%)
45+	405	102 (25%)	303 (75%)	22 (5%)	383 (95%)	357 (88%)	48 (12%)	47 (12%)	358 (88%)
Educação*									
LS	254	148 (58%)	106 (42%)	70 (28%)	184 (72%)	231 (92%)	23 (8%)	87 (34%)	167 (66%)
CP	270	118 (44%)	152 (56%)	51 (19%)	219 (81%)	244 (90%)	26 (10%)	70 (26%)	200 (74%)
SD	503	405 (81%)	98 (19%)	230 (46%)	273 (54%)	470 (93%)	33 (7%)	253 (50%)	250 (50%)
Gr	478	202 (42%)	276 (58%)	100 (21%)	378 (79%)	444 (93%)	34 (7%)	123 (26%)	355 (74%)
PG	372	118 (32%)	254 (68%)	62 (17%)	310 (83%)	353 (95%)	19 (5%)	76 (20%)	296 (80%)
País									
UK	1.044	307 (29%)	737 (71%)	159 (15%)	885 (85%)	969 (93%)	75 (7%)	162 (16%)	882 (84%)
EUA	551	482 (87%)	69 (13%)	242 (44%)	309 (56%)	513 (93%)	38 (7%)	314 (57%)	237 (43%)
Outros	282	202 (72%)	80 (28%)	112 (40%)	170 (60%)	260 (92%)	22 (8%)	133 (47%)	149 (53%)
Etnia									
Brancos	1.715	912 (53%)	803 (47%)	466 (27%)	1.249 (73%)	1.600 (93%)	115 (7%)	552 (32%)	1.163 (68%)
Negros	33	8 (24%)	25 (76%)	3 (9%)	30 (91%)	25 (76%)	8 (24%)	5 (15%)	28 (85%)
Outros	129	71 (55%)	58 (45%)	44 (34%)	85 (66%)	117 (91%)	12 (9%)	52 (40%)	77 (60%)

*Codificação: LS - Deixou a escola; CP - curso profissionalizante; SD - sem diploma; Gr - graduação; PG - diploma mestrado ou doutorado.

Tabela 1.39: Frequências das variáveis relacionadas aos traços de personalidade (Scores) nas amostras de interesse.

		Cannabis		Ecstasy		Álcool		Estimulantes	
	Amostra	Usuário	Não Usuário	Usuário	Não Usuário	Usuário	Não Usuário	Usuário	Não Usuário
	n = 1.885	991 (53%)	886 (47%)	513 (27%)	1.364 (73%)	1.742 (93%)	135 (7%)	609 (32%)	1.268 (68%)
NScore									
Muito baixo	864	404 (47%)	460 (53%)	221 (26%)	643 (74%)	803 (93%)	61 (7%)	226 (26%)	638 (74%)
Baixo	655	351 (54%)	304 (46%)	174 (27%)	481 (73%)	611 (93%)	44 (7%)	224 (34%)	431 (66%)
Médio	321	215 (67%)	106 (33%)	108 (34%)	213 (66%)	293 (91%)	28 (9%)	144 (45%)	177 (55%)
Alto	37	21 (57%)	16 (43%)	10 (27%)	27 (73%)	35 (95%)	2 (5%)	15 (41%)	22 (59%)
EScore									
Muito baixo	428	254 (60%)	174 (40%)	111 (26%)	317 (74%)	396 (93%)	32 (7%)	160 (37%)	268 (63%)
Baixo	1.011	499 (49%)	512 (51%)	254 (25%)	757 (75%)	932 (92%)	79 (8%)	296 (29%)	715 (71%)
Médio	424	227 (54%)	197 (46%)	141 (33%)	283 (67%)	400 (94%)	24 (6%)	145 (34%)	279 (66%)
Alto	14	11 (79%)	3 (21%)	7 (50%)	7 (50%)	14 (100%)	0 (0%)	8 (57%)	6 (43%)
OScore									
Muito baixo	100	19 (19%)	81 (81%)	13 (13%)	87 (87%)	93 (93%)	7 (7%)	19 (19%)	81 (81%)
Baixo	657	232 (35%)	425 (65%)	102 (16%)	555 (84%)	604 (92%)	53 (8%)	146 (22%)	511 (78%)
Médio	947	597 (63%)	350 (37%)	313 (33%)	634 (67%)	884 (93%)	63 (7%)	352 (37%)	595 (63%)
Alto	173	143 (83%)	30 (17%)	85 (49%)	88 (51%)	161 (93%)	12 (7%)	92 (53%)	81 (47%)
AScore									
Muito baixo	188	119 (63%)	69 (37%)	67 (36%)	121 (64%)	178 (94%)	10 (6%)	99 (53%)	89 (47%)
Baixo	907	509 (56%)	398 (44%)	267 (29%)	640 (71%)	842 (93%)	65 (7%)	315 (35%)	592 (65%)
Médio	736	346 (47%)	390 (53%)	172 (23%)	564 (77%)	681 (93%)	55 (7%)	187 (25%)	549 (75%)
Alto	46	17 (37%)	29 (63%)	7 (15%)	39 (85%)	41 (89%)	5 (11%)	8 (17%)	38 (83%)
CScore									
Muito baixo	319	237 (74%)	82 (26%)	124 (39%)	195 (61%)	300 (94%)	19 (6%)	166 (52%)	153 (48%)
Baixo	866	505 (58%)	361 (42%)	264 (30%)	602 (70%)	809 (93%)	57 (7%)	303 (35%)	563 (65%)
Médio	667	237 (36%)	430 (64%)	119 (18%)	548 (82%)	609 (91%)	58 (9%)	131 (20%)	536 (80%)
Alto	25	12 (48%)	13 (52%)	6 (24%)	19 (76%)	24 (96%)	1 (4%)	9 (36%)	16 (64%)

REFERÊNCIAS BIBLIOGRÁFICAS

Araujo, E.A., de Montreuil Carmona, C.U., 2007. Desenvolvimento de modelos credit scoring com abordagem de regressão logística para a gestão da inadimplência de uma instituição de microcrédito. URL: https://www.redalyc.org/pdf/1970/197014735006. pdf.

Bellini, T., Brock, E., Soares, M., Weber, A., 2020. Projeto final do curso de machine learning e modelagem estatística ofertado durante o curso de verão 2020. URL: https://github.com/taisbellini/ml _modelagem/tree/master/projeto final.

Faria, L., 2014. As 5 grandes dimensões da personalidade. URL: https://meucerebro.com/as-5-grandes-dimensoes-da-personalidade/.

Fehrman, E., Mirkes, E.M., Egan, V., 2015a. Drug consumption (quantified) data set. URL: https://archive.ics.uci.edu/ml/datasets/Drug+consumption+quantified.

Fehrman, E., Muhammad, A.K., Mirkes, E.M., Egan, V., Gorban, A.N., 2015b. The five factor model of personality and evaluation of drug consumption risk. URL: https://arxiv.org/pdf/1506.06297.pdf, arXiv:1506.06297.

Therneau, T.M., Atkinson, E.J., 2019. An introduction to recursive partitioning using the rpart routines. https://cran.r-project.org/web/packages/rpart/vignettes/ longintro.pdf.

IDENTIFICAÇÃO DE PULSARES UTILIZANDO TÉCNICAS DE APRENDIZAGEM DE MÁQUINA E MODELAGEM ESTATÍSTICA

Artur Mattia Ongarato
Programa de Pós-Graduação em Estatística - UFRGS

Eduardo Cavalli Lacerda
Programa de Pós-Graduação em Estatística - UFRGS

Felipe Grillo Pinheiro†
Programa de Pós-Graduação em Estatística - UFRGS

Matheus Daniel Pierozan

William Cechin Guarienti
ARPAC Brasil

RESUMO: A identificação de pulsares é possibilitada através da análise de padrões de emissão de radiação eletromagnética. Todavia, a maior parte das detecções de tais padrões de radiação eletromagnética consistem apenas em interferência de radiofrequência (RFI) e ruído, de modo a tornar a identificação dos sinais legítimos um grande desafio. Este trabalho propõe uma análise comparativa do emprego de algoritmos de machine learning para a classificação de candidatos como pulsares. Para tanto, foi utilizado o banco de dados HTRU2 disponibilizado pela Universidade da Califórnia em Irvine.
PALAVRAS-CHAVE: Classificação; machine learning; validação cruzada; pulsares

2.1 INTRODUÇÃO

Pulsares são estrelas raras de nêutrons que produzem emissões de rádio detectáveis na Terra, havendo interesse científico em seu estudo. De fato, é possível estudar a evolução estelar, a natureza da gravitação e a composição do meio interestelar através da observação de pulsares.

À medida que um pulsar gira, ele produz um padrão detectável de emissão de rádio em banda larga. A busca por pulsares então envolve a procura por sinais de rádio periódicos com grandes telescópios, o que se torna ainda mais desafiador ao se considerar que cada pulsar produz um padrão de emissão ligeiramente diferente a cada rotação. Assim, a detecção de um sinal potencial, chamada de "candidato" é calculada sobre muitas rotações do pulsar, durante a duração de uma observação. Sem informações adicionais, cada candidato poderia descrever um pulsar real. Porém, quase todas as detecções são causadas por interferência de radiofrequência (RFI) e ruído, dificultando a detecção de sinais legítimos.

Até recentemente, as análises de seleção de candidatos eram realizadas manualmente. Porém, essa tarefa consome muito tempo e introduz erro humano aos resultados. Assim, tem-se usado técnicas de aprendizagem de máquina para segregar de modo automático e rápido os pulsares dos demais candidatos. Portanto, tem-se então um problema de classificação binária, onde os pulsares são uma classe minoritária (positiva) e os elementos que não são pulsares são uma classe majoritária (negativa).

O estudo das técnicas de aprendizagem de máquina que será apresentado a seguir fez uso do banco HTRU2, disponível no repositório de *machine learning* da UCI. Este banco descreve uma amostra de 17898 candidatos a pulsar coletados durante a High Time Resolution Universe Survey, sendo 16259 exemplos espúrios (causados por RFI/ruído) e 1639 pulsares reais, conferidos por anotadores humanos.

Para cada observação, foram registrados dados de oito variáveis contínuas, das quais as quatro primeiras são estatísticas simples do perfil integrado do padrão de emissão e as quatro seguintes são obtidas da curva DM-SNR (medida de dispersão e razão sinal-ruído). A nona e última informação associada à cada observação do banco de dados é uma variável binária que traduz a classificação de um candidato como pulsar ou não. Assim, as variáveis disponíveis para a análise desenvolvida neste trabalho são as seguintes:

V1. Média do perfil integrado (mIP);

V2. Desvio padrão do perfil integrado (seIP);

V3. Excesso de curtose perfil integrado (ekIP);

V4. Assimetria do perfil integrado (sIP).

V5. Média da curva DM-SNR (mDM);

V6. Desvio padrão da curva DM-SNR (seDM);

V7. Excesso de curtose da curva DM-SNR (ekDM);

V8. Assimetria do DM-SNR (sDM).

V9. Classificação do candidato como 0 (a observação não se trata de um pulsar) ou 1 (a observação consiste em um pulsar).

Ainda que uma descrição pormenorizada do significado físico das variáveis supracitadas esteja fora do escopo deste trabalho, é fundamental desenvolver alguma intuição a respeito do problema em questão. Desse modo, o desenvolvimento deste trabalho iniciou-se pelo desenvolvimento de uma revisao bibliográfica, sendo que o leitor interessado nos aspectos astrofísicos desta matéria é doravante convidado a consultar Lyon et al. (2016) para maiores detalhes.

Posteriormente, o desenvolvimento do trabalho continua com a descrição da metodologia associada ao emprego das técnicas de aprendizado de máquina que foram escolhidas para compor este trabalho, a saber: Regressão Logística, *Support Vector Machines* (SVM), Redes Neurais, Árvores de Decisão e *Ensembles*. Por fim, os resultados obtidos com o desenvolvimento deste trabalho são apresentados em conjunto com as conclusões.

2.1.1 Objetivo

O objetivo deste trabalho foi comparar o desempenho de diferentes técnicas de classificação amplamente empregadas em *Machine Learning* no problema de identificação de pulsares, considerando a influência da Análise de Componentes Principais sobre as predições de tais modelos.

2.2 REVISÃO BIBLIOGRÁFICA

Diversas técnicas de aprendizagem de máquina foram empregadas para abordar o problema de identificação de pulsares. A fim de organizar a discussão, cada técnica será discutida individualmente nas subseções a seguir, sendo que os pormenores relacionados ao emprego das técnicas e a análise comparativa dos resultados são reservada às Seções 2.3 e 2.4, respectivamente.

2.2.1 Regressão Logística

A regressão logística é um método estatístico utilizado quando o objetivo é verificar a relação entre uma variável resposta binária Y e variáveis explicativas de interesse $X_1, \cdots X_n$. O modelo associado é o modelo linear generalizado definido por

$$P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = g(\boldsymbol{x})$$

$$g(\boldsymbol{x}) = \ln\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n,$$

onde $x = (x_1, \cdots, x_n)'$ é o vetor de observações das covariáveis $\{X_i \mid i \in \mathbb{N}, 1 \leq i \leq n\}$, $\{\beta_i \mid i \in \mathbb{N}, 1 \leq i \leq n\}$ são parâmetros a determinar e

$$\pi(\boldsymbol{x}) = \frac{1}{1 + e^{-g(\boldsymbol{x})}},$$

é a função sigmoíde, também denominada logística (James et al., 2013). Neste caso, a previsão do valor da variável resposta é confinada ao intervalo entre 0 e 1.

2.2.2 Support Vector Machine (SVM)

Um *Support Vector Machine* (SVM) é um classificador binário que procura um hiper- plano ótimo como uma função de decisão em um espaço de dimensões maiores. Em duas dimensões, esse hiperplano é uma linha que divide um plano em duas partes onde cada classe a ser predita está sobre esse mesmo plano (Gunn et al., 1998). As SVMs apresentam hiperparâmetros que podem ser ajustados para melhorar o classificador, deixando-a com um bom poder preditivo. Os hiperparâmetros de uma SVM são: o *kernel*, a margem, o parâmetro de regularização (C) e um parâemtro γ.

Conforme Lopez et al. (2018), a função *kernel* é uma função de similaridade que corresponde ao produto escalar em um espaço vetorial expandido, que basicamente define a forma como será feita a separação das classes. De acordo com a circunstância, podem ser empregadas funções lineares, não lineares, polinomiais, radiais ou sigmoidais. Definido o *kernel*, o SVM busca identificar o hiperplano que maximize as distâncias entre ele e os dados mais próximos (de qualquer classe), as quais estão estreitamente associadas ao conceito de margem.

O parâmetro de regularização (C, em alguns pacotes de programação computacional), por sua vez, indica o quanto a otimização da SVM tolera classificações errôneas. Assim, por exemplo, a otimização irá definir hiperplanos com menores margens para grandes valores de C. Finalmente, o parâmetro γ define uma espécie de raio de influência dos pontos associados aos dados de treinamento na definição do hiperplano de separação quando um *kernel* radial é utilizado. Assim, por exemplo, altos valores de γ acarretam em uma maior influência dos dados mais próximos do hiperplano sobre a definição do seu formato.

2.2.3 Redes Neurais

As redes neurais empregadas em aprendizado de máquina foram desenvolvidas baseadas no funcionamento do cérebro humano ou, mais especificamente, das suas unidades elementares, os neurônios. A partir do entendimento da estrutura e funcionamento do neurônio biológico, pesquisadores tentaram simular esse mesmo sistema em computador. O modelo matemático mais bem aceito recebeu o nome de *perceptron* cuja representação matemática é dada pela Figura 2.1.

Observando a Figura 2.1, tem-se os estímulos externos representados pelos sinais

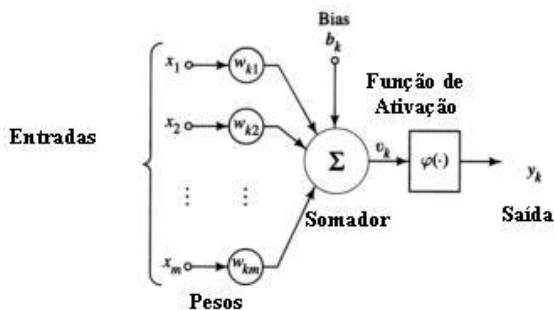



Figura 2.1: Representação simplificada do neurônio matemático.

de entrada X_i, os quais são ponderados pelos pesos W_{ki}. As entradas ponderadas, entao, são somadas a um termo constante b_k, chamado *bias*, que é passado como uma entrada de uma função de ativação $\varphi(\cdot)$. Essa função – que pode assumir diversas formas (linear,

degrau, logística, etc.), conforme discutido em Michie et al. (1994) – definirá a saída y_k, ou seja, a resposta do neurônio.

Em uma rede neural típica, a estrutura formada pelos neurônios interligados uns aos outros pode ser representada como um grafo orientado, onde cada vértice representa um neurônio e as arestas representam os pesos. Há diversas topologias de redes neurais, como aquela ilustrada na Figura 2.2, denominada de redes alimentada adiante (*feed-forwards networks*). Conforme sugere o seu nome, a informação nessa rede entrará pela camada de entrada e sairá pela última camada seguindo o fluxo de informação indicado na Figura 2.2.

Para o ajuste dos parâmetros da rede neural, o principal algoritmo é o de retropropagação do erro, baseado na regra de aprendizagem por erro, o qual pode ser pensado em termos de dois passos. No primeiro passo, os dados de entrada são passados para os neurônios de cada camada em sequência, até a camada de saída, produzindo uma resposta na rede. No segundo passo, o valor do erro de predição é transmitido no sentido inverso, ajustando os pesos, dada uma certa regra, da camada de saída até a primeira camada.

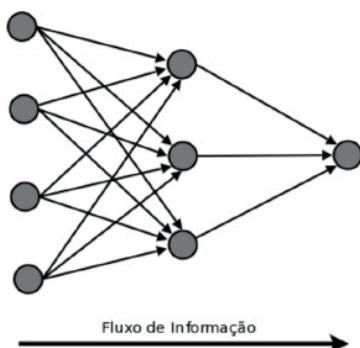



Figura 2.2: Estrutura de uma rede neural alimentada adiante.

2.2 Árvores de Decisão

Árvores de decisão são modelos de aprendizagem de máquina supervisionados bastante utilizados para resolver problemas de classificação e regressão, que se valem da estratégia de *dividir para conquistar*. Conforme ilustra a Figura 2.3, em uma árvore de Decisão, cada nó de Decisão contém um teste para algum atributo; cada folha está associada a uma classe e cada percurso da árvore, da raiz à folha, corresponde a uma regra de classificação. Desse modo, conforme elabora Gama (2004), cada folha corresponde a um hiper-retângulo no espaço definido pelos atributos, onde a interseção destes é vazia e a união é todo o espaço.

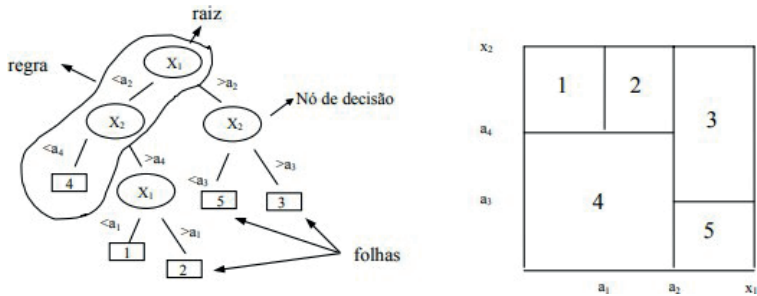



Figura 2.3: Exemplo de uma árvore de decisão e da sua respectiva representação no espaço definido pelos atributos.

De uma forma geral, a construção de uma árvore de decisão envolve a escolha dos atributos, a definição dos testes dos nós e a poda da árvore. Para tanto, a estratégia mais simples utilizada na construção de árvores de Decisão é a abordagem *Top-Down Introduction of Decision Tree* (TDIDT), também conhecida como *top-down*. É a base para muitos algoritmos bem conhecidos como o ID3, C4.5 e CART.

O TDIDT constrói a árvore através de sucessivas divisões das observações de acordo com os valores de seus atributos preditivos. É um algoritmo recursivo de busca gulosa que procura, em um conjunto de atributos, aquelas que melhor dividem o conjunto de observações em subconjuntos.

A construção de uma árvore de decisão é dada colocando inicialmente todos as observações em um nó raiz. A cada passo, é escolhido o atributo que *melhor* divide o conjunto de amostras, promovendo o crescimento da árvore. O crescimento é interrompido quando uma expansão ulterior da árvore não contribui para *melhor* explicar os dados. Evidentemente, as métricas relacionadas à definição da melhor subdivisão e ao critério de parada (ou poda, no jargão de aprendizagem de máquina) dependem das especificidades de cada algoritmo utilizado.

2.2.5 Ensembles

Ensembles, também conhecidos como máquinas de comitê, são estruturas que também seguem a estratégia de dividir para conquistar. De modo mais específico, são constituídos por um conjunto de estimadores individuais organizados de forma paralela, cuja saída decorre de alguma combinação particular dos estimadores individuais (Friedman et al., 2001).

Existem diversas razões que motivam a utilização de *ensembles*, de caráter estatístico, computacional e representacional. Do ponto de vista estatístico, por exemplo, quando os dados são insuficientes, o algoritmo pode realizar uma média das decisões dos componentes individuais do *ensemble*, reduzindo a chance de fazer a escolha por uma hipótese errada.

Na construção de *ensembles*, existem dois aspectos fundamentais a serem considerados: a definição da base dos estimadores e o método utilizado para a combinação deles. Do ponto de vista da definição da base dos estimadores, é fundamental sopesar a acurácia individual contra a diversidade entre eles, de modo que o conjunto de classificadores não possua apenas alta acurácia, mas também que seus erros sejam descorrelacionados (Kuncheva, 2004). No que diz respeito ao método de combinação, é preciso determinar quais são os estimadores que terão suas respostas utilizadas para compor a saída do algoritmo. Dentre os métodos mais importantes encontrados na literatura que implementam *ensembles*, destacam-se os algoritmos de *bagging* (Breiman, 1996), *random forest* (Breiman, 2001) e *boosting* (Schapire, 1990).

No *bagging*, diferentes sub-amostras com reposição do mesmo conjunto de dados são empregados para a construção de diferentes classificadores, conforme ilustra a Figura 2.4. Desse modo, objetiva-se reduzir a variância das previsões através da combinação das saídas dos classificadores. O *random forest*, por sua vez, é um algoritmo baseado em *bagging*, que utiliza múltiplas Árvores para predizer a variável resultado. Contudo, há uma pequena diferença: ao construir cada árvore de Decisão a partir dos dados, é empregado apenas um subconjunto aleatório de variáveis preditoras, enquanto no *bagging* todas as preditoras são utilizadas.

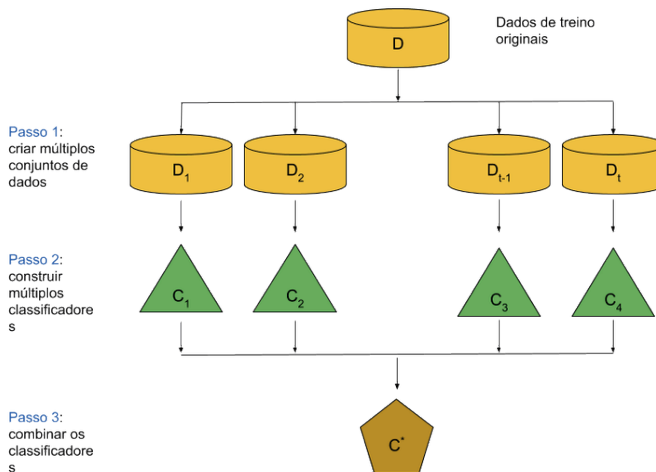



Figura 2.4: Ilustração do algoritmo de *bagging*

Finalmente, no *boosting* cada classificador é treinado usando um conjunto de treinamento diferente, do mesmo modo que anteriormente. A principal distinção em relação ao *bagging* é que os conjuntos de dados reamostrados são construídos especificamente para gerar aprendizados complementares, e a importância do voto é ponderado com base no desempenho de cada modelo, em vez da atribuição de mesmo peso para todos os votos. Com efeito, no *boosting*, os modelos não são mais treinados de forma independente entre

si, mas de forma sequencial, de modo que os novos modelos enfoquem nas observações que os modelos anteriores tiveram mais dificuldade de classificar.

2.3 METODOLOGIA

Para avaliar a capacidade preditiva na classificação dos pulsares foram utilizadas as técnicas classificação mencionadas na seção anterior. Cada técnica possui suas particularidades e desdobramentos em termos dos parâmetros utilizados nas respectivas implementações computacionais na linguagem de programação R. Nesta seção serão apresentados uma breve análise descritiva dos dados (Seção 2.3.1), os detalhamentos dos cenários em que os modelos foram avaliados (Seção 2.3.2), as técnicas empregadas na classificação dos pulsares (Seção 2.3.3) e as estatísticas utilizadas para comparação dos modelos (Seção 2.3.4).

2.3.1 Análise descritiva dos dados

Nesta etapa, buscou-se fundamentalmente investigar as relações existentes entre as variáveis preditoras e entre as variáveis preditoras e a variável resposta. Assim, inicialmente construiu-se a matriz de correlação amostrais, de modo a quantificar as correlações entre as variáveis preditoras do modelo e contribuir no ganho de intuição sobre os resultados. Para o banco de dados em questão, a matriz de correlação pode ser ilustrada com uso da Figura 2.5.

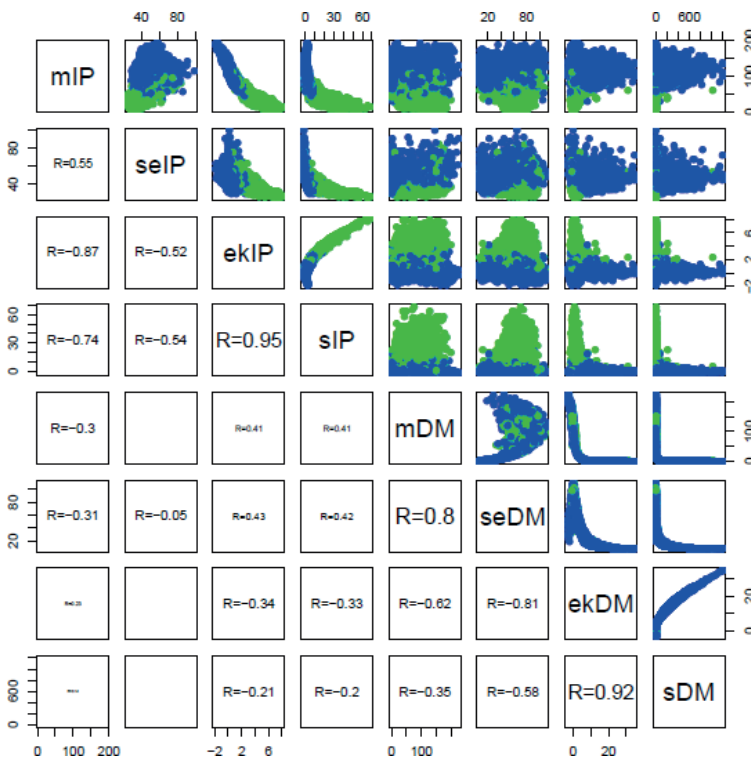



Figura 2.5: Matriz de gráficos de dispersão e correlações amostrais

A Figura 2.6 apresenta o *box-plot* dos dados referentes ao banco original enquanto a Figura 2.7 apresenta o *box-plot* dos dados referentes ao banco PCA. A Tabela 2.1 apresenta as correlações entre cada variável preditora (mIP, seIP, ekIP, sIP, mDM, seDM, ekDM e sDM), e a variável resposta (classe do pulsar).

Tabela 2.1: Matriz de correlação entre as variáveis preditoras e a variável resposta (Y = classe do pulsar).

	mIP	seIP	ekIP	sIP	mDM	seDM	ekDM	sDM
Y	-0,67	-0,36	0,79	0,7	0,4	0,49	-0,39	-0,25

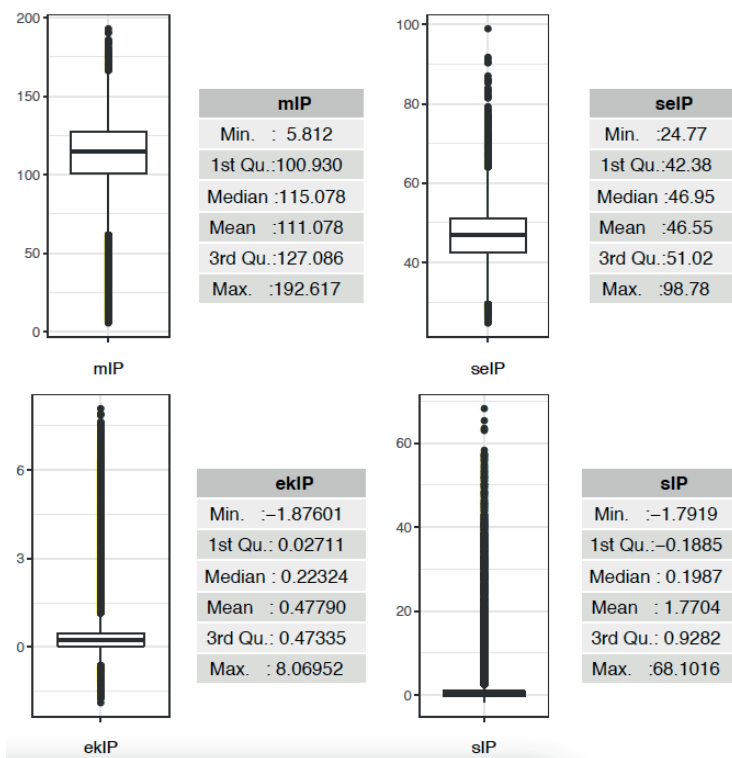



Figura 2.6: Box-plot das variáveis referentes ao Perfil Intergrado (mIP, seIP, ekIP e sIP)

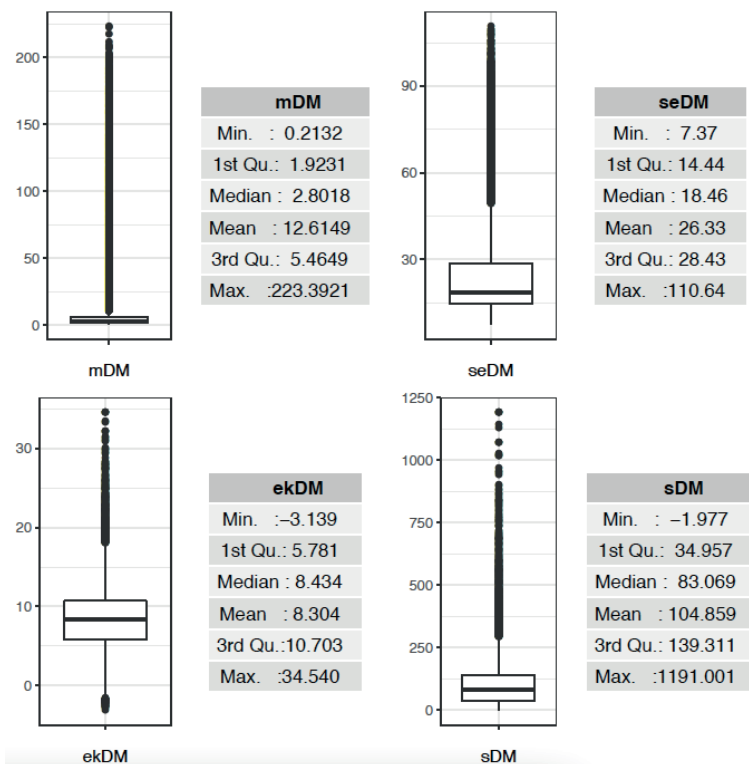



Figura 2.7: Box-plot das variáveis referentes às medidas de dispersão e razão sinal-ruído (mDM, seDM, ekDM e sDM)

A Análise de Componentes Principais (PCA) realizada nas covariáveis contínuas do banco de dados indica que as duas primeiras componentes principais explicam aproximadamente 78,48% da variabilidade dos dados. A decomposição da variabilidade total pode ser visualizada na Tabela 2.2.

Tabela 2.2: Decomposição da variabilidade total em termos das componentes principais.

Componente	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Proporção da variância	0,517	0,268	0,101	0,057	0,032	0,020	0,003	0,002
Proporção cumulativa	0,517	0,785	0,886	0,943	0,976	0,995	0,998	1,000

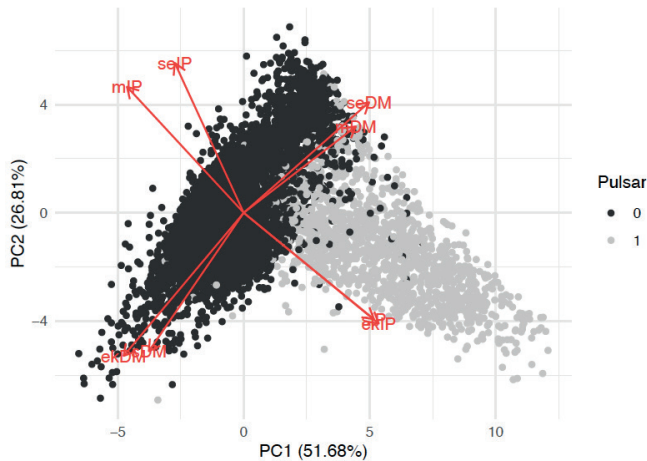



Figura 2.8: Dispersão dos dados no espaço gerado pelas duas componentes principais

As covariáveis mais correlacionadas com a primeira componente, PC1, foram ekIP, sIP e seDM. Já para a segunda componente, PC2, temos as covariáveis seIP, ekDM e sDM como as mais influentes. Tais relações podem ser visualizadas na Figura 2.8 que mostra a dispersão dos dados nas componentes PC1 e PC2 e a direção das variáveis no PCA. Finalmente, complementando a caracterização dos dados, a Figura 2.9 mostra a dispersão dos pulsares positivos e negativos nas componentes principais par a par.

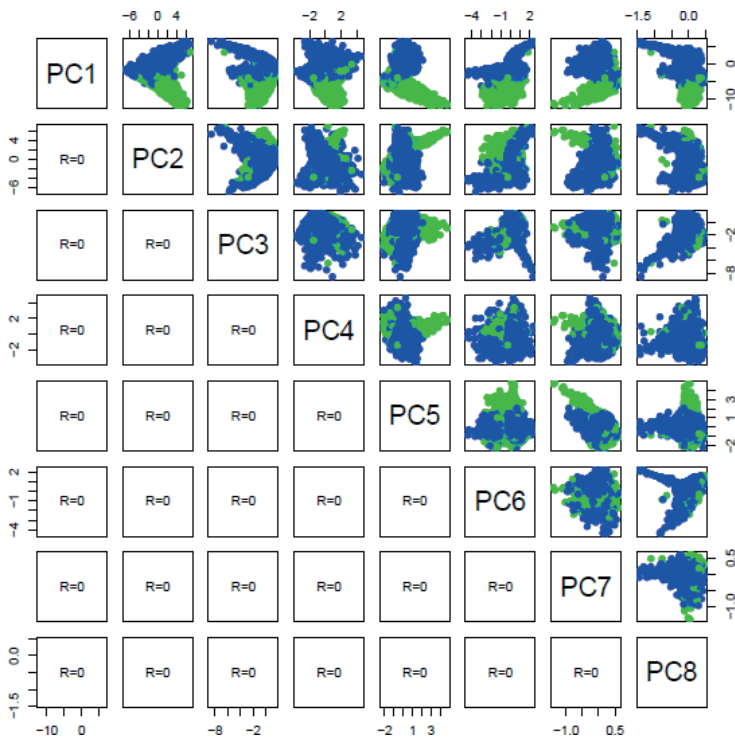



Figura 2.9: Matriz de gráficos de dispersão e correlações amostrais

2.3.2 Cenários nas avaliações dos modelos

Com a finalidade de avaliarmos a potencial contribuição da Análise de Componentes Principais (PCA) sobre o desempenho das técnicas de classificação, os modelos de cada técnica foram ajustados tanto para os dados originais do banco quanto para os dados correspondentes às oito componentes principais calculadas com base no banco inteiro. Neste caso, o PCA foi aplicado como uma espécie de transformação dos dados. Os bancos de dados foram subdivididos em duas partes, treino e teste, nas proporçoes de 0,75 e 0,25, respectivamente. Em cada técnica, os modelos foram ajustados para a amostra de treino e, a partir deste modelo, foram elaboradas a matriz de confusão e as estatísticas pertinentes associadas ao banco de treino e de teste. Todas as análises foram realizadas utilizando a linguagem de programação R.

2.3.3 Técnicas de Classificação

Regressão Logística

Na regressão logística foi ajustado um único modelo utilizando a função train do pacote caret (Kuhn, 2020) e escolhendo o método glm. Como parâmetros de pré- transformação dos dados utilizou-se center e scale. Nos parâmetros de controle utilizou-se o método cv (validação cruzada) e 10 como o número de *folds*.

Support Vector Machine (SVM)

Para avaliar a capacidade de classificação dos pulsares por meio da técnica *Support Vector Machine* (SVM) foi utilizado o pacote e1071 (Meyer et al., 2019) do R. Dois modelos foram ajustados com a amostra de teste: um para o *kernel* linear; outro, para o radial. Com o *kernel* linear utilizou-se como parâmetros o cost = 10 e a opção de padronização, scale = TRUE. Para o ajuste com o *kernel* radial foi realizado um pré-teste com a função tune.svm, também do pacote e1071, para a escolha dos parâmetros gamma e cost. Os parâmetros ótimos foram gamma = 0,8 e cost = 2,9. As observações receberam diferentes pesos de acordo com a classe do pulsar. Pulsares negativos (não-pulsares) receberam peso w1 = 6,15, enquanto os positivos, peso w2 = 61,12. Os valores dos pesos foram sendo ajustados conforme foram sendo realizados os testes, de modo a aumentar a relevância da classe minoritária. A opção de padronização foi também scale = TRUE. As matrizes de confusão para os dados de treino e teste foram obtidas – assim como nos demais casos – por meio da função confusionMatrix, do pacote caret.

Redes neurais

A técnica de redes neurais foi aplicada em um único modelo ajustado por meio da função train, do pacote caret utilizando o método nnet. Foi utilizada a função tune.nnet, do pacote e1071, para a escolha dos valores ótimos para os parâmetros size e decay. Como parâmetros de pré-transformação dos dados utilizou-se center e scale. Nos parâmetros de controle utilizou-se o método cv e 10 como o número de *folds*. Em tuneGrid os valores ótimos para parâmetros size e decay, derivados do tune.nnet, foram 14 e 0,26, respectivamente.

Árvores de Decisão

Com a técnica de Árvores de decisão foram avaliados dois modelos utilizando diferentes pacotes no R. O primeiro modelo foi ajustado por meio da função *rpart*, do pacote rpart (Therneau and Atkinson, 2019), com os parâmetros minsplit = 5, minbucket = 3 e com o critério split = information, e os demais parâmetros escolhidos como os padrões da função. No segundo modelo, utilizou-se a função *train* do pacote caret, com parâmetro tuneLength= 10, que é ligado à granulosidade do parâmetro de *tuning* e parâmetros de controle: method = cv e number = 10.

Ensembles

Para as máquinas de comitê, foram empregados os algoritmos *bagging*, *random forest* e *boosting*, os quais já foram introduzidos na Seção 2.2. Para o *bagging*, foram criadas árvores de Decisão usando a função bagging do pacote ipred (Peters and Hothorn, 2019). Neste modelo os parâmetros de controle foram: nbagg = 100, coob = TRUE, minsplit = 20, minbucket = 2 e cp = 0. O algoritmo utilizado para a classificação através de *random forest*, foi aquele da função randomForest do pacote homônimo randomForest (Liaw and Wiener, 2002), com parâmetros de ajuste ntree = 500, mtry = 6, importance = TRUE. O parâmetro mtry = 6 foi escolhido via pré-teste com um pequeno *grid* de variação de valores. Por fim, o algoritmo de *boosting* utilizado foi aquele implementado pela função gbm do pacote homônimo gbm (Greenwell et al., 2019), com parâmetros distribution = multinomial, n.trees = 1000, shrinkage = 0.01, n.minobsinnode = 2 e interaction. depth = 4.

2.3.4 Estatísticas na avaliação dos modelos

Para descrevermos as estatísticas utilizadas para avaliar o desempenho dos métodos, deve-se observar que os quatro possíveis desfechos de classificação que compõem a matriz de confusão são:

- Verdadeiro positivo (VP): total de positivos corretamente previstos;
- Verdadeiro negativo (VN): total de negativos corretamente previstos;
- Falso positivo (FP): total de negativos erroneamente previstos;
- Falso negativo (FN): total de positivos erroneamente previstos.

Com estas medidas é possível calcular uma série de estatísticas que retratam o desempenho das técnicas de classificação em Aprendizado de Máquina. A Tabela 2.3 resume as métricas utilizadas para avaliar a performance dos métodos de classificações para o desfecho binário e raro (pulsar). De acordo com Lyon et al. (2016), avaliar o desempenho do classificador em dados desequilibrados é difícil. Por exemplo, no banco de dados utilizado neste trabalho, um método que classificasse todas as observações como não-pulsares teria uma alta acurácia (igual a 0,9084, mais especificamente), apesar desta classificação não ter utilidade alguma na prática. Desta forma, no caso de conjuntos de dados grandes e desequilibrados, a média geométrica (*G-Mean*), é uma medida adequada, pois descreve a proporção de acurácia positiva e negativa independentemente do tamanho da classe, uma vez que cada parte de sua fórmula leva em conta apenas o percentual de classificação correta dentro das classes positiva e negativa. Desta forma, esta estatística é sensível ao desbalanço entre sensibilidade e especificidade. Além da *G-Mean*, costuma-se usar também o *F-Score*. O *F-Score* depende da precisão e da sensibilidade, sendo portanto influenciado pelo desequilíbrio entre falsos positivos e falsos negativos.

Estatística	Definição
Acurácia $$\frac{VP+VN}{VP+FP+FN+VN}$$	Proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo. Esta medida é altamente suscetível a desbalanceamentos do conjunto de dados e pode facilmente induzir a uma conclusão errada sobre o desempenho do sistema.
Taxa de falsos positivos (FPR) $$\frac{FP}{FP+VN}$$	Proporção de falsos positivos em relação ao total de observações negativas no banco ($FP + VN$). Representa a capacidade do sistema em predizer erroneamente a existência da condição para casos que realmente não as têm.
Precisão $$\frac{VP}{VP+FP}$$	Proporção dos verdadeiros positivos em relação ao total de classificados como positivos ($VP + FP$). Representa a capacidade do sistema em predizer corretamente a presença da condição para casos que a classificação foi positiva.
Sensibilidade $$\frac{VP}{VP+FN}$$	Proporção de verdadeiros positivos em relação ao total de observações positivas no banco ($VP + FN$). Representa a capacidade do sistema em predizer corretamente a presença da condição para casos que realmente as têm.

Especificidade	$\dfrac{VN}{VN+FP}$	Proporção de verdadeiros negativos em relação ao total de observações negativas no banco (*FP* + *V N*). Representa a capacidade do sistema em predizer corretamente a ausência da condição para casos que realmente não as têm.
G-Mean	$\sqrt{\dfrac{VP}{VP+FN}\dfrac{VN}{VN+FP}}$	Proporção de acurácia positiva e negativa independentemente do tamanho da classe. É uma medida sensível ao desbalanço entre sensibilidade e especificidade.
F-Score	$\dfrac{2\left(\dfrac{VP}{VP+FP}\dfrac{VP}{VP+FN}\right)}{\dfrac{VP}{VP+FP}+\dfrac{VP}{VP+FN}}$	Depende da precisão e da sensibilidade, portanto é influenciado pelo desequilíbrio entre o número de falsos positivos e falsos negativos.

Tabela 2.3: Estatísticas utilizadas para avaliação da performance dos métodos de classificação.

2.4 RESULTADOS E CONCLUSÕES

Nesta seção, serão comparados os resultados das diferentes técnicas de classificação aplicadas aos dados, assim como a influência da Análise de Componentes Principais. Para isso, calculou-se as estatísticas descritas na Seção 2.3.4 para os resultados obtidos com as técnicas empregadas, conforme ilustram as Figuras 2.10 e 2.11 no Apêndice. O gráfico de Sensibilidade foi repetido propositalmente nas Figuras 2.10 e 2.11 com o objetivo de facilitar as análises. De especial interesse, no entanto, são as estatísticas *G-Mean* e *F-Score*, devido ao desfecho raro do problema abordado.

Adicionalmente foi realizada uma análise dos significância dos parâmetros obtidos para a Regressão Logística, tanto para o banco original, quanto para o banco PCA. A Tabela 2.4 apresenta os resultados para os dados originais, onde é possível constatar que as variáveis ekDM e sDM não foram significativas, tendo em vista que os p-valores das mesmas foram acima de $0,05$.

Tabela 2.4: Significância dos parâmetros obtidos com a Regressão Logística para os dados nas dimensões originais

Variáveis	Coeficiente	Pr(>l z l)	Significância
(Intercept)	−432, 269	< 10−4	***
mIP	0, 8474	< 10−4	***
seIP	−0, 3011	< 10−4	***
ekIP	717, 698	< 10−4	***
sIP	−389, 845	< 10−4	***
mDM	−0, 7837	< 10−4	***
seDM	0, 9720	< 10−4	***
ekDM	0, 1870	0, 67	NS
sDM	−0, 5197	0, 16	NS

Por outro lado, realizando a mesma análise para o banco PCA observa-se que todas as variáveis foram significativas. Os resultados da análise de significância procedida é apresentada na Tabela 2.5.

Tabela 2.5: Significância dos parâmetros obtidos com a Regressão Logística para os dados nas dimensoes do PCA

Variáveis	Coeficiente	Pr(>l z l)	Significância
(Intercept)	–432,269	< 10–4	***
PC1	–278,714	< 10–4	***
PC2	0,9900	< 10–4	***
PC3	0,2507	0,003	***
PC4	0,7482	< 10–4	***
PC5	–0,4200	< 10–4	***
PC6	0,6147	< 10–4	***
PC7	111,070	< 10–4	***
PC8	0,1476	0,043	*

2.4.1 Acurácia

Todos os métodos apresentaram acurácia acima de 0,97, tanto no teste quanto no treino, para os banco original e para os banco PCA (Figura 2.10). Chama a atenção o fato de *AD Random Forest* ter obtido acurácia unitária (ajuste perfeito) na amostra de treino. Como consequência, as demais estatísticas envolvendo esta técnica também apresentaram valores máximos, à exceção da Taxa de Falsos Positivos, que, neste caso foi 0. Para evitar repetição, não mais falaremos de seu desempenho em relação às amostras de treino. Destaca-se também que, para os banco original e para o banco PCA, *AD Boosting* e *Random Forest* obtiveram, respectivamente, as maiores acurácias na amostra de teste. Além disso, o *SVM (kernel radial)* teve o pior desempenho na amostra de teste, para ambos os bancos. No banco PCA, houve um aumento da acurácia na amostra de treino, mas uma queda na amostra de treino. Esta técnica parece sofrer bastante influência de uma transformação com a PCA. Contudo, conforme citado anteriormente, a acurácia sozinha pode não ser uma medida adequada para a avaliação dos modelos e técnicas, quando o desfecho é um evento raro.

2.4.2 Especificidade

Exceto pelo *SVM (kernel radial)*, todos os métodos obtiveram especificidade bastante elevada (acima de 0,99), tanto no treino, quanto no teste, no banco original e no banco PCA. Como será verificado adiante, isso se deve ao fato de a sensibilidade do *SVM (kernel radial)* ter sido maior que a dos demais métodos; e, em geral, aumentar a sensibilidade

faz com que a especificidade diminua. Cabe observar que o *SVM (kernel linear)* foi o mais específico na amostra de teste em ambos os tipos de banco.

2.4.3 Sensibilidade

Analisando a sensibilidade podemos entender por que a especificidade do *SVM (kernel radial)* é visualmente menor que a dos demais no banco de dados original. Esta foi a técnica que obteve maior sensibilidade na amostra de teste (0,868 contra 0,856 do *AD Bagging*, o segundo melhor). Entretanto, para os dados PCA, seu desempenho em sensibilidade foi o pior entre todos os modelos.

2.4.4 Precisão

Assim como a especificidade, a precisão leva em conta os falsos positivos em seu denominador. Na amostra de teste, tanto para os dados originais quanto para os dados PCA, o *SVM (kernel linear)* apresentou o melhor desempenho, seguido pela Regressão Logística (Figura 2.11). Já o *SVM (kernel radial)* foi o que obteve pior resultado.

2.4.5 Taxa de falsos positivos

Nesta estatística, os métodos que mais se aproximaram de 0 (quanto menor, melhor) na amostra de teste, foram novamente o *SVM (kernel linear)* e a Regressão Logística. Outra vez o *SVM (kernel radial)* teve o pior desempenho entre os métodos.

2.4.6 G-Mean

Ao avaliar a *G-Mean*, devemos novamente lembrar que ela é uma ponderação entre sensibilidade e especificidade. Dito isso, é interessante notar que, na amostra de teste com os dados originais, o *SVM (kernel radial)*, que teve pior precisão, taxa de falsos positivos e até mesmo especificidade, obteve, juntamente com o *AD Bagging*, a maior *G-Mean* entre os métodos implementados. Isso se deve ao fato de sua sensibilidade ser a mais elevada. Por outro lado, novamente na amostra de teste com os dados originais, o *SVM (kernel linear)*, que apresentou a maior especificidade, foi o que teve menor *G-Mean*, justamente por esta medida ser sensível ao desbalanço entre a grande especificidade e a menor sensibilidade. Já para os dados PCA, o *SVM (kernel radial)* e a árvore de decisão simples (AD Simples) obtiveram desempenho inferior na amostra de teste, onde as redes neurais tiveram o melhor resultado.

2.4.7 F-Score

Nos dados originais, *AD Boosting* e *AD Bagging* dominaram na amostra de teste. Nos dados PCA, *AD Random Forest* foi levemente superior aos dois citados anteriormente. Por

sua vez, o *SVM (kernel radial)* apresentou o pior *F-Score* (apesar de ter o terceiro melhor desempenho na amostra de treino com os dados originais e o segundo, com os dados PCA).

2.4.8 CONSIDERAÇÕES FINAIS

Analisando conjuntamente as estatísticas, em especial nas amostras de teste, pois o interesse é prever se uma observação é ou não pulsar, destacamos as técnicas *SVM (kernel radial)* e *AD Bagging*, que obtiveram o maior *G-Mean*. Ao compararmos o *F- Score* de ambas, *AD Bagging* aparenta superioridade. Isto deve-se ao fato de o *F-Score* ser sensível à precisão, estatística na qual o *SVM (kernel radial)* teve o pior desempenho entre as técnicas utilizadas. Apesar disso, este último foi o método que apresentou maior sensibilidade. Em outras palavras, *SVM (kernel radial)* identificou mais os verdadeiros pulsares, às custas de classificar, por outro lado, mais falsos positivos.

Ao realizar este trabalho, concluiu-se que, para um problema de classificação genérico, sobre o qual pouco se conhece *a priori*, é recomendável que sejam testadas diferentes técnicas e ajustados diversos modelos em cada uma, pois não há uma técnica unanimamente melhor. É necessário conhecer o contexto dos dados, uma vez que é importante entender os custos relativos à precisão e sensibilidade. De fato, pode-se gastar muitos recursos tendo que revisar um falso positivo, assim como pode-se perder uma informação de grande valia com um falso negativo.

APÊNDICE: FIGURAS

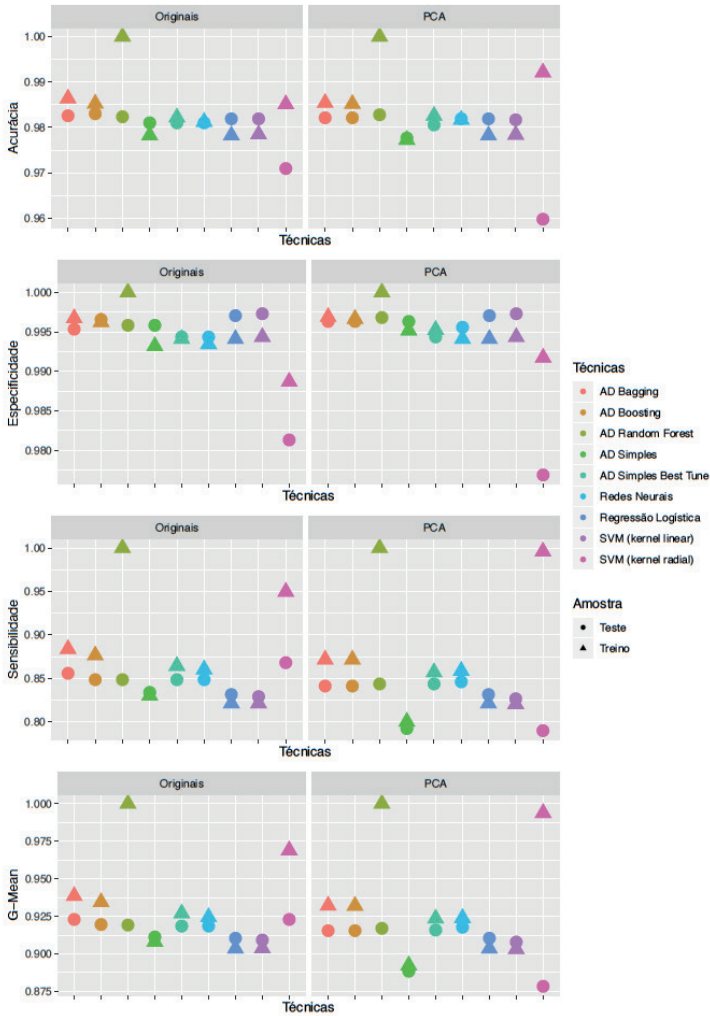



Figura 2.10: Acurácia, Especificidade, Sensibilidade e *G-Mean* da predição de classificação dos pulsares em amostras de treino e teste, utilizando os dados originais e os dados transformados via PCA de acordo com diferentes técnicas de classificação empregadas em *Machine Learning*.

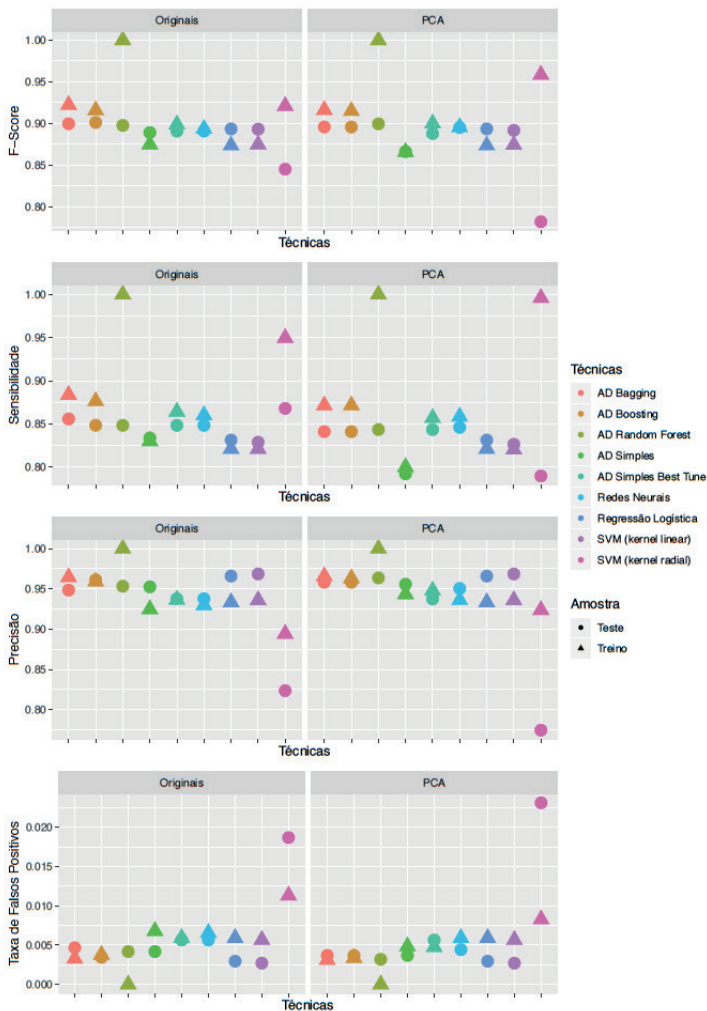



Figura 2.11: *F-Score*, Sensibilidade, Precisão e Taxa de Falsos Positivos da predição de classificação dos pulsares em amostras de treino e teste, utilizando os dados originais e os dados transformados via PCA de acordo com diferentes técnicas de classificação empregadas em *Machine Learning*.

REFERÊNCIAS BIBLIOGRÁFICAS

Breiman, L., 1996. Bagging predictors. Machine learning 24, 123–140.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning. 1 ed., Springer series in statistics New York.

Gama, J., 2004. Árvores de decisão. https://www.cin.ufpe.br/~if684/EC/aulas/Aula-arvores-decisao-SI.pdf. Accessed: 2020-02-15.

Greenwell, B., Boehmke, B., Cunningham, J., Developers, G., 2019. gbm: Generalized Boosted Regression Models. URL: https://CRAN.R-project.org/package=gbm. r package version 2.1.5.

Gunn, S.R., et al., 1998. Support vector machines for classification and regression. ISIS technical report 14, 5–16.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning: with Applications in R. Springer Texts in Statistics, Springer New York.

Kuhn, M., 2020. caret: Classification and Regression Training. URL: https://CRAN. R-project.org/package=caret. r package version 6.0-85.

Kuncheva, L.I., 2004. Classifier ensembles for changing environments, in: International Workshop on Multiple Classifier Systems, Springer. pp. 1–15.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2, 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

Lopez, M.A., Sanz, I.J., Lobato, A.G., 2018. Aprendizado de máquina em plataformas de processamento distribuído de fluxo: Análise e detecção de ameaças em tempo real. Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)- Minicursos .

Lyon, R.J., Stappers, B., Cooper, S., Brooke, J., Knowles, J., 2016. Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. Monthly Notices of the Royal Astronomical Society 459, 1104–1123.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2019. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. URL: https://CRAN.R-project.org/package=e1071. r package version 1.7-3.

Michie, D., Spiegelhalter, D.J., Taylor, C., et al., 1994. Machine learning. Neural and Statistical Classification 13, 1–298.

Peters, A., Hothorn, T., 2019. ipred: Improved Predictors. URL: https://CRAN. R-project.org/package=ipred. r package version 0.9-9.

Schapire, R.E., 1990. The strength of weak learnability. Machine learning 5, 197–227.

Therneau, T.M., Atkinson, E.J., 2019. An introduction to recursive partitioning using the rpart routines. https://cran.r-project.org/web/packages/rpart/vignettes/ longintro.pdf.

PANORAMA DA ENERGIA E DESENVOLVIMENTO SUSTENTÁVEL MUNDIAL: ANÁLISE DE AGRUPAMENTO

Jéssica Duarte
Escola de Engenharia - UFRGS

Jonas Tieppo
Escola de Engenharia - UFRGS

Luiz Fernando Bez
Instituto de Matemática e Estatística - UFRGS

Lara Werncke Vieira
Escola de Engenharia - UFRGS

RESUMO: Como parte da agenda de desenvolvimento para o ano de 2030, a Assembleia Geral da ONU aprovou o que chamou de um "projeto para atingir um futuro melhor e mais sustentável para todos". Esse projeto se formalizou nos objetivos para o Desenvolvimento Sustentável (SDG's - Sustainable Development Goals), totalizando 17 objetivos. Neste trabalho nos propomos a analisar o SDG 7 - Energia Limpa e Acessível, utilizando técnicas de agrupamento, com o objetivo de identificar grupos de países com comportamento semelhante em relação ao objetivo em questão.
PALAVRAS-CHAVE: *k*-means, *k*-NN, modelo hierárquico, desenvolvimento sustentável

3.1 INTRODUÇÃO

No ano de 2015, como parte da agenda de desenvolvimento para o ano de 2030, a Assembleia Geral da ONU aprovou o que chamou de um "projeto para atingir um futuro melhor e mais sustentável para todos". Esse projeto se formalizou nos objetivos para o Desenvolvimento Sustentável (SDG's - Sustainable Development Goals). Os 17 objetivos globais definidos são:

1. Acabar com a pobreza em todas as suas formas, em todos os lugares
2. Acabar com a fome, alcançar a segurança alimentar e melhoria da nutrição e promover a agricultura sustentável
3. Assegurar uma vida saudável e promover o bem-estar para todos, em todas as idades
4. Assegurar a educação inclusiva, equitativa e de qualidade, e promover oportunidades de aprendizagem ao longo da vida para todos
5. Alcançar a igualdade de gênero e empoderar todas as mulheres e meninas
6. Assegurar a disponibilidade e gestão sustentável da água e saneamento para todos.

7. Assegurar o acesso confiável, sustentável, moderno e a preço acessível à energia para todos
8. Promover o crescimento econômico sustentado, inclusivo e sustentável, emprego pleno e produtivo e trabalho decente para todos
9. Construir infraestruturas resilientes, promover a industrialização inclusiva e sustentável e fomentar a inovação
10. Reduzir a desigualdade dentro dos países e entre eles
11. Tornar as cidades e os assentamentos humanos inclusivos, seguros, resilientes e sustentáveis
12. Assegurar padrões de produção e de consumo sustentáveis
13. Tomar medidas urgentes para combater a mudança climática e seus impactos
14. Conservação e uso sustentável dos oceanos, dos mares e dos recursos marinhos para o desenvolvimento sustentável
15. Proteger, recuperar e promover o uso sustentável dos ecossistemas terrestres, gerir de forma sustentável as florestas, combater a desertificação, deter e reverter a degradação da terra e deter a perda de biodiversidade
16. Promover sociedades pacíficas e inclusivas para o desenvolvimento sustentável, proporcionar o acesso à justiça para todos e construir instituições eficazes, responsáveis e inclusivas em todos os níveis
17. Fortalecer os meios de implementação e revitalizar a parceria global para o desenvolvimento sustentável

Esses objetivos são bastante amplos, e de várias formas interdependentes - alguns objetivos apresentando uma complementaridade enquanto outros se prejudicam mutuamente, de acordo com Lusseau and Mancini (2019). As maiores críticas aos objetivos são justamente seu aspecto generalista, seu número elevado, e a competitividade entre objetivos. Apesar das críticas, os SDG's foram um resultado da assembleia geral que não foi criticado por nenhuma grande ONG, pelo contrário, recebendo largo suporte das mesmas.

Num esforço de medir o progresso mundial nesses objetivos, e ajudar que os mesmos sejam cumpridos, foram criados 169 objetivos específicos para estes 17 SDG's. Cada um desses 169 objetivos específicos têm entre 1 e três indicadores, totalizando 232 indicadores oficialmente aprovados pela ONU. Ainda com o objetivo de adicionar clareza ao processo, a ONU disponibiliza bases de dados com informações à respeito desses indicadores e objetivos específicos.

Tendo acesso às bases de dados desses SDG's, surgiram as seguintes questões: como os países se distribuem no que diz respeito à esses objetivos? Seria possível notar um agrupamento dos mesmos baseado no seu progresso em algum dos SDG's? Mais especificamente, neste trabalho nos propomos a analisar o SDG 7 - Energia Limpa e Acessível. Para isso serão recuperados os dados dos indicadores, os mesmos serão tratados e adequados para a análise, e por fim, diversas técnicas de agrupamento serão aplicadas.

3.1.1 Sobre o SDG 7

Objetivos para 2030 miram no aumento do acesso à energia enquanto que a fração de energia renovável do mercado também aumenta. Isso envolve, dentre outros, o aumento de eficiência energética, cooperação internacional para a proliferação de tecnologias de energia renovável, e mais investimento em infraestrutura para energia limpa. Ainda, o SDG pede atenção particular para os países menos desenvolvidos. Como um exemplo, em 2017, 57% da população global depende primariamente de energia limpa para cozinhar, estando longe dos 95% almejados.

3.2 BASE DE DADOS

Os dados dos indicadores são disponibilizados na página oficial da ONU. Contudo, sua formatação é difícil de trabalhar. Esses dados também estão disponíveis no The World Bank (2019), onde podem ser obtidos em um formato muito mais direto para análise. Para cada indicador é possível que existam dados dos últimos trinta anos de 263 países/regiões disponibilizados na base de dados. Contudo, existe uma porcentagem grande de dados faltantes para alguns países. Além disso, alguns indicadores apresentam pouca informação para muitos anos atrás (os SDG's não estavam estabelecidos e poucos países têm essas informações), bem como muitos países ainda não disponibilizaram os dados de muitos indicadores para os últimos dois ou três anos.

Os parâmetros nesta base de dados escolhidos por representarem os indicadores do SDG 7 são apresentados abaixo.

1. Acesso a combustíveis limpos e tecnologias para cozinhar (% da população) - *Access to clean fuels and technologies for cooking*,
2. Acesso a eletricidade (% da população) - *Access to electricity*
3. Produção de eletricidade a partir de fontes renováveis, excluindo hidrelétricas (% do total) - *Energy production from renewable sources, excluding hydroelectric*,
4. Nível de intensidade energética da energia primária (MJ/US 2011 PIB PPP) - *Energy intensity level of primary energy*
5. Consumo de energia renovável (% do consumo total final de energia) - *Renewable energy consumption*
6. Produção de eletricidade renovável (% da produção total de eletricidade) - *Renewable electricity output*
7. Uso de energia (kg de petróleo equivalente per capita) - *Energy use*

Em seguida, outra escolha feita nesta análise foi a da fatia temporal a ser estudada. Gostaríamos de fazer uma análise de agrupamento o mais recente possível, mas também gostaríamos de ter o maior número de dados possível. Sendo assim, escolhemos fazer a análise para dados referentes ao ano de 2015 como um bom trade-off.

Mesmo escolhendo 2015 como ano referência, ainda existem alguns casos de missing data. De fato, veremos que um dos indicadores, devido ao excesso de dados

faltantes, foi excluído da análise, e por isso foram adicionados dois outros indicadores, da mesma base de dados, que representam uma quantidade semelhante. Quantificamos a proporção de dados faltantes para cada um dos parâmetros:

1. Acesso a combustíveis limpos e tecnologias para cozinhar (% da população): 9,6%
2. Acesso a eletricidade (% da população): 0%
3. Produção de eletricidade a partir de fontes renováveis, excluindo hidrelétricas (% do total): 48,1%
4. Nível de intensidade energética da energia primária (MJ / US 2011 PIB PPP): 8,7%
5. Consumo de energia renovável (% do consumo total final de energia): 4,8%
6. Produção de eletricidade renovável (% da produção total de eletricidade): 14,8%
7. Uso de energia (kg de petróleo equivalente per capita): 32,6%

Na Figura 3.1 podemos ver a distribuição de dados faltantes para cada parâmetro ou indicador do banco de dados. O indicador 3, "Produção de eletricidade a partir de fontes renováveis, excluindo hidrelétricas (% do total)", possui mais da metade dos dados faltantes (48,1%), e portanto foi excluído da análise.

Com relação ao tratamento de dados para os dados faltantes, substituiu-se os valores faltantes pela média do indicador. As médias foram calculadas por grupo de países de acordo com a sua região no globo, tendo em vista que a média do banco de dados não seria representativa.

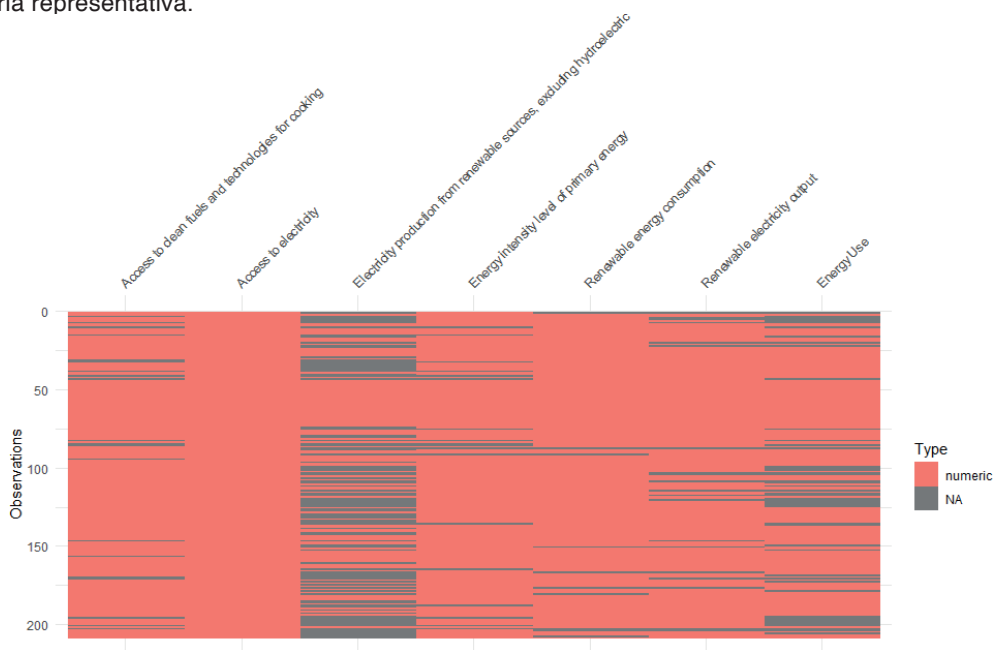



Figura 3.1: Distribuição de dados faltantes para cada parâmetro.

As regiões consideradas foram as seguintes: Mundo Árabe, Pequenos países Caribenhos, Leste Asiático e Pacífico (excluindo alta renda), Europa e Ásia Central (excluindo alta renda), América Latina e Caribe (excluindo alta renda), Oriente Médio e África do Norte (excluindo alta renda), Pequenos países das Ilhas do Pacífico, Sul da Ásia, África Subsariana (excluindo alta renda).

3.3 ANÁLISE DOS COMPONENTES PRINCIPAIS

O primeiro passo da análise foi a aplicação do método de análise de componentes principais (PCA) aos dados, a fim de melhorar a visualização e verificar se é possível explicar a sua variabilidade reduzindo dimensionalidade, de forma visual.

A Tabela 3.1 mostra o quanto cada componente principal (PC) representa de variabilidade no espaço *p-dimensional*, assim como sua homogeneidade (desvio padrão). É importante aqui analisar que a primeira componente do PCA teve uma proporção de variância de 50.87%. Em outras palavras, ela pode representar um pouco mais de 50% do padrão dos dados, com um desvio padrão de 1,74. Com duas componentes, pode-se chegar a explicar 70,62 % da variância, o que permite afirmar, até este ponto, que com duas componentes pode-se ter um ótima representação dos dados disponíveis. A partir da obtenção dos componentes principais, é possível analisar a relação com as variáveis iniciais, representada pela Tabela 3.2.

Tabela 3.1: Representação das componentes principais na variabilidade dos dados.

	PC1	PC2	PC3	PC4	PC5	PC6
Desvio padrão	1,7470	1,0887	0,9648	0,7081	0,4258	0,3866
Proporcão de variância	0,5087	0,1975	0,1551	0,0836	0,0302	0,0249
Proporcão acumulada	0,5087	0,7062	0,8613	0,9449	0,9751	1,0000

Tabela 3.2: Pesos (loadings vectors) de cada componente do PCA.

	PC1	PC2	PC3	PC4	PC5	PC6
Acesso a combustíveis limpos e tecnologias para cozinhar	-0,51	0,15	-0,26	0,21	-0,37	-0,69
Acesso a eletricidade	0,51	0,02	-0,26	0,30	-0,26	0,71
Nível de intensidade energética da energia primária	0,22	0,69	0,33	0,61	0,04	0,02
Consumo de energia renovável	0,52	0,12	-0,19	-0,18	-0,80	0,08
Produção de eletricidade renovável	0,29	0,18	-0,85	0,12	0,38	-0,02
Uso de energia	-0,28	0,67	0,00	-0,67	0,08	0,10

A Tabela 3.2 mostra os pesos (ou loadings) obtidos para cada componente principal. A contribuição é dada pelo módulo do peso, logo seu sinal não afeta a análise. Tendo em vista este conceito, pode-se concluir que as protagonistas para a primeira componente principal são três: “Acesso a combustíveis limpos e tecnologias para cozinhar”, “Acesso a eletricidade” e “Consumo de energia renovável”. Na segunda componente, outras duas variáveis exercem o papel principal, sendo elas: “Nível de intensidade energética da energia primária e “Produção de eletricidade renovável” e “Uso de energia”. O fato das variáveis com maior peso não serem iguais nas duas componentes é esperado pelo método.

No método PCA, busca-se a cada componente o conjunto de dados que formam uma base ortogonal com o a componente anterior. Então, as variáveis protagonistas na PC2 tendem a não ser as mesmas que as do PC1. Além disso, pode-se dizer que as variáveis “Acesso a combustíveis limpos e tecnologias para cozinhar”, “Acesso a eletricidade” e “Consumo de energia renovável” estão correlacionadas de certa forma entre si. Logo, as mesmas, segundo o conjunto de dados, possuem uma correlação *menor* (mas não fraca, note-se que a “Produção de eletricidade renovável” tem uma contribuição significativa na PC1) com as outras variáveis.

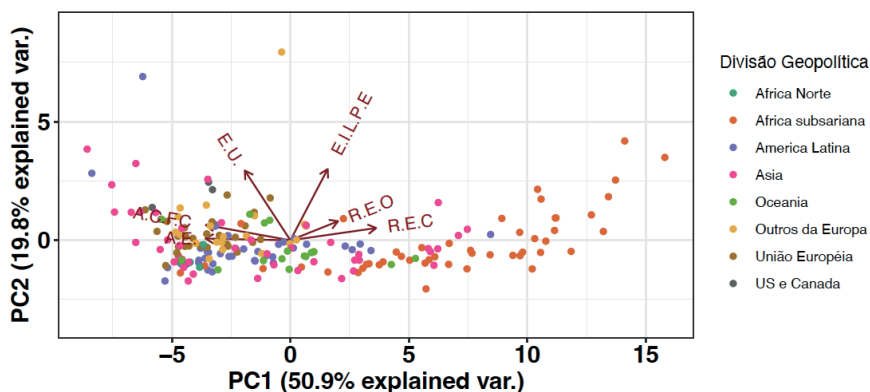
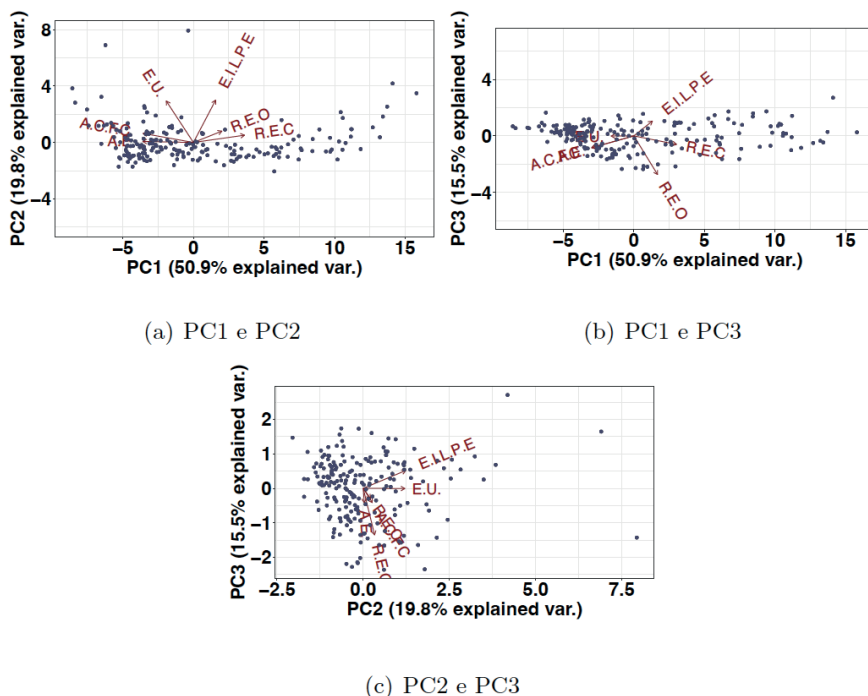



Figura 3.2: Dispersão dos dados na dimensao de PC1 e PC2, salientando-se a divisão geopolítica proposta.

As mesmas conclusões podem ser observadas a partir da visualização gráfica, apresentada a Figura 3.2. Vê-se que as variáveis “Acesso a combustíveis limpos e tecnologias para cozinhar”, “Acesso a eletricidade” e “Consumo de energia renovável” estão bem próximas no eixo horizontal (PC1). Graficamente, isso representa uma correlação entre as variáveis, ou seja, se o acesso de um país à eletricidade cresce, o acesso à combustíveis limpos também cresce. Já a “Nível de intensidade energética da energia primária”, é menos correlacionada com essas duas variáveis. Em outras palavras, se a intensidade de energia primária de um país é grande, não necessariamente este apresenta um grande acesso a combustíveis renováveis.

Conforme o esperado, as duas principais separações entre os grupos se devem justamente aos parâmetros com mais influência nas componentes. A Figura 3.3(b) mostra mais claramente esta divisão, em que pode-se observar os vetores “Acesso a combustíveis limpos e tecnologias para cozinhar”, “Acesso a eletricidade” e “Consumo de energia renovável” na direção de maior variabilidade dos dados e oposta a direção do vetor “Produção de eletricidade renovável”.

(a) PC1 e PC2

(b) PC1 e PC3

(c) PC2 e PC3

Figura 3.3: Dispersão dos dados nas dimensões de (a) PC1 e PC2; (b) PC1 e PC3 e (c) PC2 e PC3.

3.4 AGRUPAMENTO *K* -MEANS

A primeira técnica de agrupamento utilizada foi a técnica denominada *k*-means. Esta técnica agrupa os dados em *k* grupos, onde cada dado, representado por um ponto no plano, pertencerá ao grupo de centro mais próximo (métrica euclidiana) do ponto. O número de clusters ou grupos desejado deve ser informado pelo usuário.

Devido à distribuição dos dados observada na dimensão dos PC’s, foi decidido aplicar o *k*-means na dimensão dos PC’s ao invés das variáveis originais. O número total de grupos selecionados foi igual a 6, para que posteriormente os grupos possam ser comparados com os continentes do globo. Agrupamentos em menos grupos apresentavam um extenso número de países por grupo e, por outro lado, agrupamentos em mais grupos se tornava excessivo.

A Figura 3.4 possibilita uma primeira visualização, na dimensão PC1 e PC2, de acordo com o resultado da distribuição por *k*-means. A Figura 3.5 apresenta uma visualização no gráfico nas dimensões PC1, PC2 e PC3.

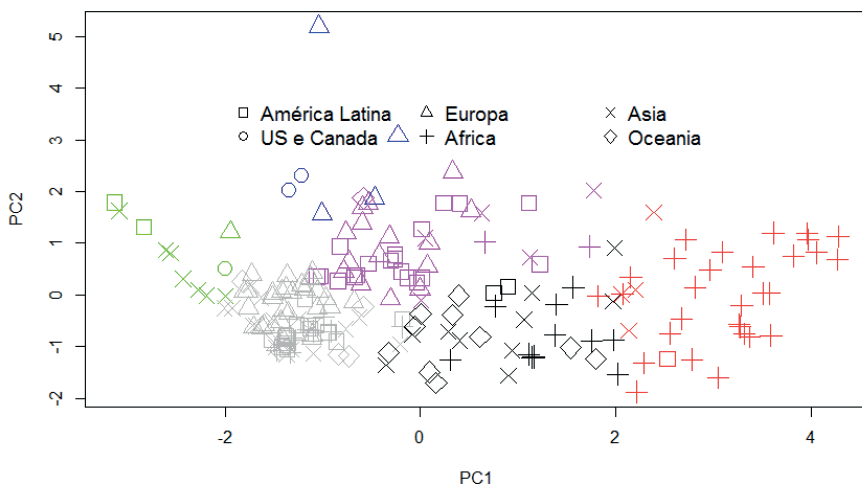



Figura 3.4: *k*-means nas dimensões de PC1 e PC2.

Inicialmente a aplicação do método PCA já demonstrou uma tendência geral dos dados dividida por dois grandes aspectos, conforme foi apresentado na Fig.3.3(a).

(a) Vista (a)

(b) Vista (b)

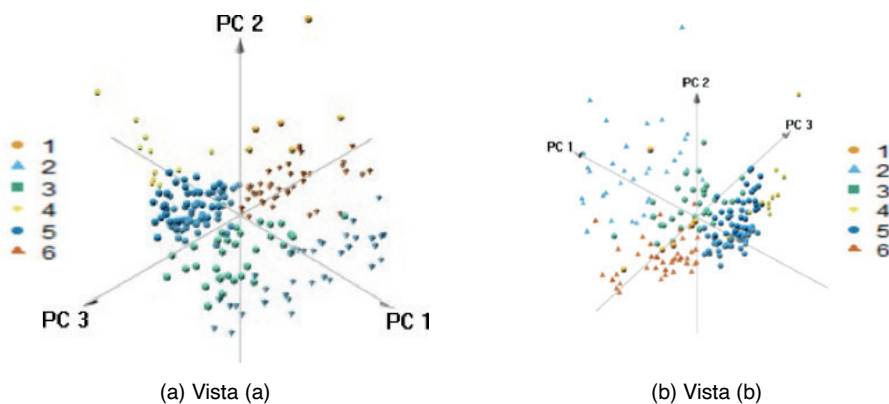



Figura 3.5: Agrupamento *k*-means 3D.

Aprofundando a análise com a aplicação de *k*-means pode-se perceber duas divisoes claras entre os grupos, sendo elas o nível de desenvolvimento do país e a sua matriz elétrica voltada para fontes renováveis de geração.

O nível de desenvolvimento do país está diretamente relacionado ao seu acesso a energia, visto que um país desenvolvido é um país com alto percentual de acesso a energia para toda a população e com alto consumo. Por outro lado, o perfil de geração de energia do país está diretamente relacionado ao seu consumo e geração de energia elétrica a partir de fontes renováveis.

Nas tabelas em apêndice são apresentados os dados de cada país que compõe os 6 agrupamentos. A Tabela 3.3 apresenta a média e o desvio padrão dos parâmetros analisados por grupo. A enumeração dos grupos é definida pelo método *k*-means. Para fins de observação dos dados, os grupos foram ordenados na ordem decrescente do parâmetro "Acesso a combustíveis limpos e tecnologias para cozinhar", representando o nível de desenvolvimento geral dos grupos.

Conforme Tabela 3.3, podemos analisar as médias e desvios padrões dos grupos por indicador. Dos 6 grupos obtidos, o grupo 1 é composto por países desenvolvidos e com matriz elétrica predominantemente renovável, com acesso a energia limpa para cozimento e acesso a eletricidade iguais a 100%. Da mesma forma, os índices de energia renovável são elevados comparados aos outros grupos. Por outro lado, o grupo 4 é composto por países desenvolvidos, mas de matriz não renovável. Desta forma, os indicadores de consumo de energia e produção de energia elétrica renovável são os menores entre os agrupamentos. Os grupos 5 e 6 são grupos medianos, divididos pelos indicadores de energia renovável. O grupo 5 apresenta valores baixos, enquanto o grupo 6 apresenta valores mais altos, ficando atrás apenas do grupo 1. Os grupos 3 e 2 são grupos de países não desenvolvidos energeticamente, com baixo acesso a energia limpa para cozimento e a eletricidade. Ainda assim, a diferenciação entre os dois grupos se dá pelo acesso a energia elétrica, que é extremamente baixo para o grupo 2.

Tabela 3.3: Média e desvio padrão (entre parênteses) dos parâmetros, para os grupos determinados pelo método de agrupamento *k*-means.

Grupo	Nº de países	ACLTC	AE	NIEEP	CER	PER	UE
1	6	100,00 (0,00)	100,00 (0,00)	7,44 (4,26)	44,81 (21,04)	74,96 (19,96)	8.527,41 (3.864,77)
4	11	97,61 (3,92)	100,00 (0,00)	6,69 (4,29)	3,03 (3,21)	6,58 (9,01)	9.872,82 (3.173,61)
5	79	91,86 (10,68)	98,45 (4,78)	5,03 (2,75)	9,19 (7,45)	13,12 (10,34)	2.212,06 (1.362,94)
6	43	83,48 (17,73)	96,36 (8,85)	4,29 (1,80)	32,37 (18,44)	65,29 (20,26)	1.625,65 (975,41)
3	34	27,71 (17,11)	75,98 (17,80)	4,75 (1,40)	35,70 (18,77)	25,92 (19,49)	805,06 (367,71)
2	35	10,32 (12,26)	33,08 (17,19)	8,29 (5,34)	75,09 (16,49)	55,77 (32,08)	571,36 (215,91)

Nota: ACLTC = Acesso a combustíveis limpos e tecnologias para cozinhar; AE = Acesso a eletricidade; NIEEP = Nível de intensidade energética da energia primária; CER = Consumo de energia renovável; PER = Produção de eletricidade renovável; UE = Uso de energia

Vale ressaltar que, no contexto do presente estudo, países desenvolvidos não estao sendo avaliados como países com o PIB elevado ou alto índice de crescimento econômico, mas sim em relação ao seu perfil energético. O agrupamento através do método *k*-means mostrou resultados coerentes e confere com o cenário real dos países. Ainda assim, uma nova avaliação foi realizada para verificar a homogeneidade do agrupamento proposto frente a agrupamentos esperados tendo em base o conhecimento geral desses países. Desta forma, comparamos o agrupamento obtido com o agrupamento dos países por continente. O resultado é apresentado na Tabela 3.4.

Tabela 3.4: Média e desvio padrão (entre parênteses) por continente.

Cont.	Nº de países	ACLTC	AE	NIEEP	CER	PER	UE
América do Norte	3	96,29 (6,43)	99,75 (0,43)	5,72 (1,33)	13,87 (5,42)	43,23 (29,64)	5.910,8 (2.541,2)
Europa	51	94,45 (9,57)	100,00 (0,00)	5,01 (2,56)	22,15 (16,25)	35,84 (27,98)	3.351,6 (2.490,9)
América Central e Sul	41	84,46 (18,76)	95,88 (9,66)	4,86 (3,37)	20,77 (20,17)	38,32 (28,21)	1.801,9 (2.924,1)
Ásia	44	67,49 (33,38)	90,53 (15,38)	5,32 (2,60)	20,53 (24,32)	26,06 (31,38)	(2.657,0) (3.285,4)
Oceania	16	46,16 (28,30)	90,72 (11,58)	5,57 (1,50)	20,58 (15,29)	20,57 (14,89)	1.706,2 (825,7)
África	53	28,97 (33,03)	48,94 (27,82)	6,71 (4,86)	56,87 (29,91)	41,31 (33,10)	772,4 (610,6)

Nota: ACLTC = Acesso a combustíveis limpos e tecnologias para cozinhar; AE = Acesso a eletricidade; NIEEP = Nível de intensidade energética da energia primária; CER = Consumo de energia renovável; PER = Produção de eletricidade renovável; UE = Uso de energia

Com o objetivo de melhorar a análise dentro de cada grupo, dois países foram selecionados por grupo como países representativos, ou seja, são países com os índices próximos às médias encontradas em cada grupo. O objetivo é situar o contexto atual do país tendo em vista desenvolvimento econômico e geração de energia elétrica e, entao, comparar com o agrupamento obtido pelo *k*-means. Os países são apresentados na Tabela 3.5.

Os indicadores de energia renovável foram essenciais para a divisão entre os grupos 1 e 4 e entre os 5 e 6. Desta forma, uma maneira de analisar o cenário real dos grupos através dos países representativos é avaliando sua matriz elétrica. Todos os dados utilizados para as matrizes elétricas por fonte são provenientes da IEA (2017). A Figura 3.6 (a)-(d) apresenta as matrizes elétricas por fonte para o Canadá, Noruega, Estados Unidos e Emirados Árabes, respectivamente.

Tabela 3.5: Países representativos por grupo.

k	Países representativos	Outros
1	Canadá, Noruega	Groelandia, Finlândia, Suécia
4	EUA, Emirados Árabes	Trinidad e Tobago, Qatar, Luxemburgo
5	México, Rússia	França, Argentina, China
6	Brasil, Portugal	Paraguai, Dinamarca, Afeganistão
3	Índia, Filipinas	Nicarágua, Camboja, Ilhas Fiji
2	Kenia, Burkina Faso	Haiti, Sudão, Nepal

Os dois grupos são compostos por países energeticamente desenvolvidos, entretanto, a divisão pela presença de energia renovável avaliando suas matrizes elétricas é clara. A Noruega apresenta matriz elétrica renovável quase em sua totalidade, com presença predominante da fonte hidrelétrica. O Canadá apresenta a matriz mais diversificada com relação as fontes de geração, mas ainda assim mais da metade da geração é proveniente de fontes renováveis. Por outro lado, os Estados Unidos e os Emirados Árabes são fortemente marcados pela presença de geração de energia a partir de fontes fósseis. Em especial, os Emirados Árabes são quase em sua totalidade caracterizados pela geração com o uso de gás natural.

Por fim, analisaremos a diferenciação entre os grupos 5 e 6 devido aos indicadores de energia renovável. As matrizes energéticas por fonte são apresentadas na Figura 3.7(a)-(d) para o México, Rússia, Brasil e Portugal, respectivamente. Conforme o esperado, os países apresentam matrizes elétricas diversificadas em comparação com os grupos anteriores. Retomando a análise para a separação entre os grupos 5 e 6, tanto o México quanto a Rússia têm alta produção de eletricidade por fontes fósseis. Por outro lado, o Brasil e Portugal se diferem pela presença de fontes renováveis. Vale ressaltar que apesar das matrizes serem apresentadas ao longo dos anos, os agrupamentos foram realizados para um cenário fixo no ano de 2015.

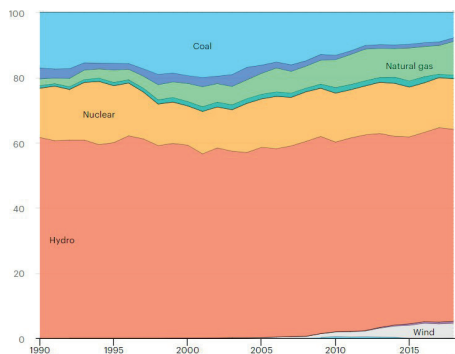
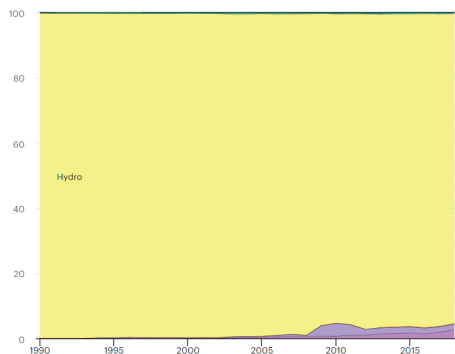
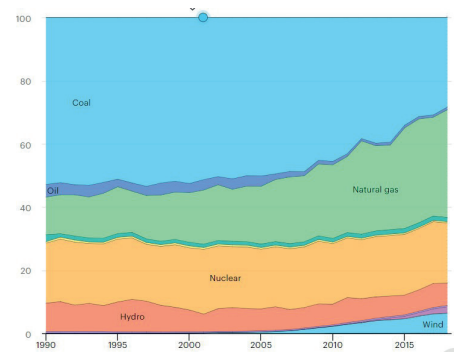



Figura 3.6: Matriz elétrica por fonte de geração no período de 1990 a 2017 para (a) Canadá; (b) Noruega; (c) Estados Unidos e (d) Emirados Árabes.

Em relação ao desenvolvimento energético, países desenvolvidos apresentam capacidade de geração de energia muito mais alta do que países subdesenvolvidos ou em desenvolvimento. Os Estados Unidos possui capacidade de geração de aproximadamente 4.000TWh, a Índia 1.500TWh e o Kenia de 10TWh.

3.5 MÉTODO HIERÁRQUICO AGLOMERATIVO

Métodos hierárquicos são métodos que envolvem a construção de uma hierarquia, que são estruturas do tipo árvore. Diferentemente do *k*-means, não precisam de uma condição inicial, o que garante certa reprodutibilidade dos resultados. O método necessita a definição de uma matriz de dissimilaridade, responsável por identificar a proximidade dos indivíduos de cada grupo. Os pares escolhidos para comporem um determinado grupo *são aqueles que possuem a menor dissimilaridade inter-grupo* (Hastie et al., 2009). Claramente o *k*-means também garante reprodutibilidade quando a semente do gerador de números aleatórios é mantida constante. O método hierárquico não tem essa necessidade, porém tem uma larga

gama de métricas e métodos de aglomeração disponíveis. A escolha dessas métricas pode ser vital para o agrupamento.

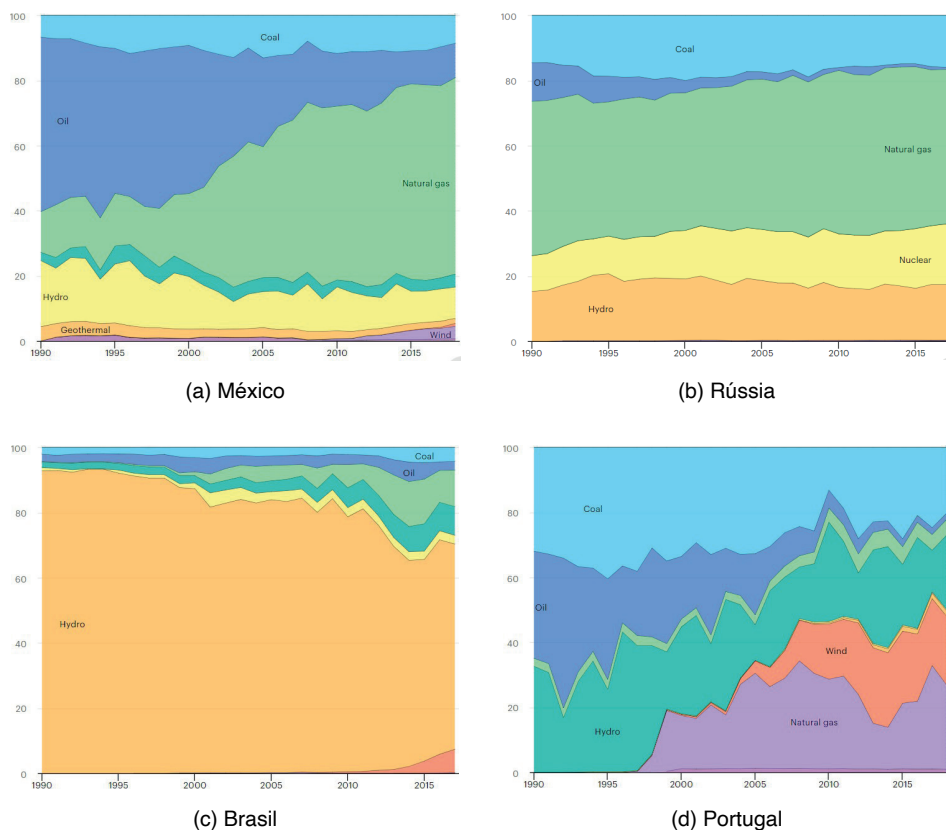
(a) México (b) Rússia

(c) Brasil (d) Portugal

Figura 3.7: Matriz elétrica por fonte de geração de 1990 a 2017 para (a) México; (b) Rússia; (c) Brasil e (d) Portugal.

A função hclust do pacote stats (R Core Team, 2019) implementa o algoritmo de agrupamento hierárquico a partir de uma matriz de dissimilaridades. A matriz de dissimilaridade é gerada a partir de um conjunto de dados x através do comando dist(x,method), onde method é a métrica a ser empregada. As seguintes métricas estão disponíveis no R: Euclidean, Manhanttan, Maximum, Canberra, Binary e Minkowski. Mais especificamente, para dois pontos $x = (x_1, \cdots, x_n)$ e $y = (y_1, \cdots, y_n)$ em $\mathbb{R}^n$, as métricas mencionadas são definidas a seguir.

Euclidean

Esta é a conhecida métrica euclideana, definida por

$$d_{eucl}(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{\frac{1}{2}}. \quad (3.1)$$

Manhattan

A métrica se baseia na distância entre dois pontos é dada pela soma absoluta de suas coordenadas, isto é

$$d_{manh}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n}|x_i - y_i|. \quad (3.2)$$

Maximum

É a máxima distância entre dois indivíduos x_i e y_i, também conhecida como norma suprema. É expressa por:

$$d_{max}(\boldsymbol{x},\boldsymbol{y}) = \max\{|x_i - y_i| : i \in \{1,\cdots,n\}\}. \quad (3.3)$$

Canberra

É uma medida numérica entre pois pontos no espaço, geralmente utilizada para a comparação de listas ordenadas, isto é

$$d_{canb}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i| + |y_i|}. \quad (3.4)$$

Ressaltamos que alguns algoritmos de aglomeração não funcionaram com essa métrica.

Minkowsky

É chamada de distância L^p, dada por

$$d_p(\boldsymbol{x},\boldsymbol{y}) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{\frac{1}{p}} \quad (3.5)$$

O método também exige a definição de um algoritmo de aglomeração. Há oito algoritmos disponíveis no pacote, denotados por Ward.D, Ward.D2, single, complete, average, mcquitty, median e centroid. A descrição dos algoritmos pode ser encontrada na documentação da função hclust do pacote stats (R Core Team, 2019).

Devido ao elevado número de métodos e métricas, foi decidido olhar como o agrupamento se comportou de maneira geral em algumas situações, para depois efetuar uma análise mais aprofundada. As figuras encontram-se na seção de anexos. Na próxima seção será apresentado apenas o agrupamento que resultou em grupos bem divididos, e também alguns contra-exemplos de clusterização, ressaltando as patologias de *chaining* e *crowding*.

3.5.1 Métrica Euclidiana e método *Ward.D2*

A combinação da métrica para dissimilaridade euclidiana com o agrupamento *ward.D2* foi o que gerou resultados preliminares melhores, vistos pela Figura 3.8. Os grupos ficaram separados, e parecem seguir uma lógica geopolítica coerente, haja vista que países com proximidades geográficas e econômicas foram na média agrupados. Foram analisados outros 39 casos de clusterização, presentes na parte dos anexos. Alguns desses também apresentaram uma boa divisão, como as Figuras 3.38 e 3.45, enquanto outras apresentaram patologias conhecidas, discutidas na próxima seção. Devido ao escopo do trabalho, apenas a Figura 3.8 será analisada mais profundamente. A Figura 3.9 mostra o dendograma resultante da análise utilizando a métrica euclidiana e agrupamento Ward.D2.

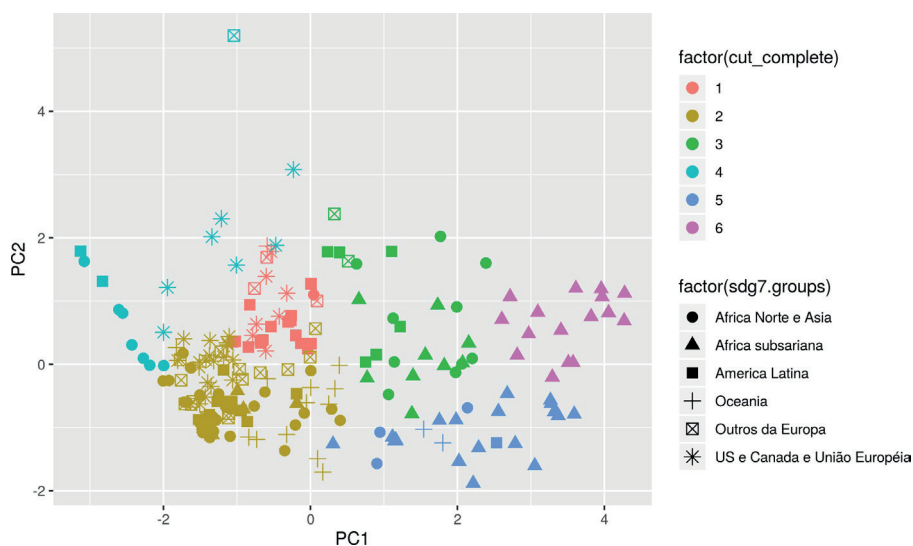



Figura 3.8: Agrupamento com o método *Ward.D2*.

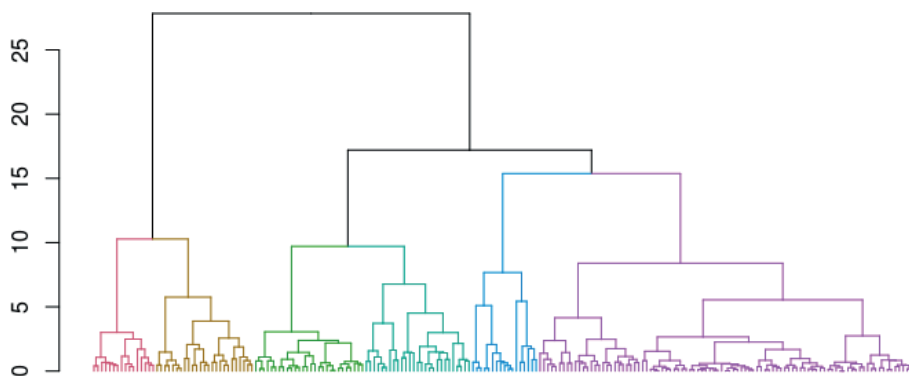



Figura 3.9: Dendograma com o método *Ward.D2*, utilizando a métrica euclidiana.

Devido a dificuldade de avaliação dos resultados, uma tabela com a quantidade de países por grupos e suas características será apresentada, como feito para o caso de *k*-means (Tabela 3.6). A composição por continente encontra-se presente na Tabela 3.7.

Tabela 3.6: Média e desvio padrão (entre parênteses) dos indicadores divididos entre os clusters.

Agrupamento	Nº de Países	ACLTC	AE (%)	NIEEP	CER (%)	PER (%)	UE (BTU per capita)
Cluster 1	28	91,92 (8,54)	98,15 (3,20)	4,02 (1,51)	23,82 (10,50)	59,96 (13,16)	1.782,2 (1.042,0)
Cluster 2	95	83,30 (22,79)	97,85 (5,46)	5,04 (2,58)	11,78 (10,70)	15,06 (11,96)	2.038,5 (1.317,5)
Cluster 3	27	45,29 (27,60)	78,32 (19,17)	4,49 (1,81)	52,21 (24,42)	68,91 (36,60)	823,9 (3.600,7)
Cluster 4	17	98,46 (3,45)	100,00 (0,00)	6,96 (4,43)	17,78 (24,42)	30,71 (36,60)	9.398,0 (3.600,7)
Cluster 5	25	11,98 (15,00)	42,49 (18,74)	6,39 (4,84)	58,73 (24,06)	16,57 (12,66)	648,5 (303,1)
Cluster 6	16	8,75 (11,09)	27,60 (13,07)	10,07 (5,00)	80,99 (11,28)	80,81 (16,72)	625,80 (239,9)

Nota: ACLTC = Acesso a combustíveis limpos e tecnologias para cozinhar; AE = Acesso a eletricidade; NIEEP = Nível de intensidade energética da energia primária; CER = Consumo de energia renovável; PER = Produção de eletricidade renovável; UE = Uso de energia

Tabela 3.7: Composição dos Clusters por Continente.

Grupo	América Latina	EUA e Canadá	União Européia	Europa (Outros)	África Subs.	Ásia	Oceania	Outros
1	57,1%	0,0%	25,0%	10,7%	0,0%	0,0%	3,6%	3,6%
2	16,8%	0,0%	18,9%	16,8%	4,2%	5,3%	24,2%	13,7%
3	22,2%	0,0%	0,0%	7,4%	33,3%	0,0%	37,0%	0,0%
4	11,8%	17,6%	23,5%	5,9%	0,0%	0,0%	41,2%	0,0%
5	4,0%	0,0%	0,0%	0,0%	76,0%	0,0%	12,0%	8,0%
6	0,0%	0,0%	0,0%	0,0%	100,0%	0,0%	0,0%	0,0%

O grupo 6 foi o que apresentou todos os seus países pertencentes à África subsariana, representado por países mais pobres, que possuem pouco acesso à eletricidade e energia limpas. O agrupamento para esse caso sucedeu de forma similar ao ocorrido para o *k*-means. O grupo 5 também é composto por países pobres, alguns da Oceania e América Latina. O grupo 4 fora o que se encaixou a maior parte dos países ricos. Os países latinos que foram agrupados podem ser vistos como um fenômeno de *crowding*.

O algoritmo não separou os países ricos em dois, da mesma maneira que o *k*-means separou. Os grupos 1 e 2 podem ser vistos como intermediários no quesito energético, que incluí países de todos os continentes, inclusive o Brasil. O Grupo 3 foi composto por países bem pobres da América Latina e alguns da África subsariana, como Namíbia, Costa Rica. O grupo, no entanto, foi o que apresentou o maior desvio padrão, talvez sinalizando um comportamento médio global.

Exemplo de *Chaining*

Vimos que alguns métodos de agrupamento hierárquicos aglomerativos podem gerar grupos com patologias. Uma delas é o chamado *chaining*, onde elementos são ligados majoritariamente ao mesmo grupo, formando um grupo muito grande e outros pouco expressivos. A Figura 3.10 é um dos exemplos de chaining encontrados na nossa análise. Ela foi feita utilizando-se o método single para linkagem dos elementos/grupos.

No apêndice se encontram outros exemplos de agrupamentos com esta patologia, nas Figuras 3.10, 3.21 e 3.28.

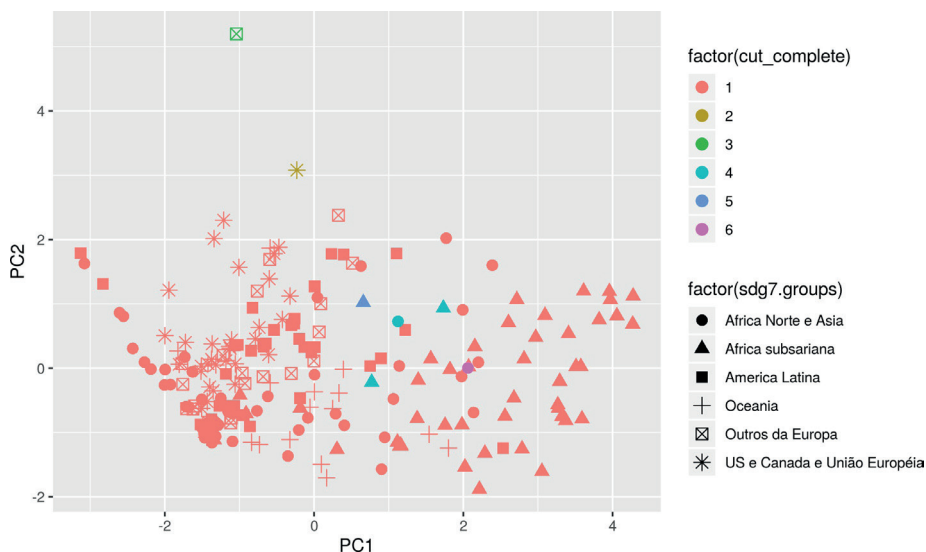



Figura 3.10: Agrupamento com o método *Single*.

Exemplo de *Crowding*

Outra problema que pode surgir em agrupamentos hierárquicos é o fenômeno de *crowding*. Na Figura 3.11 podemos ver que alguns elementos do grupo 2 estao cercados pelos grupos 4 e 5, e possivelmente deveriam pertencer aos mesmos. Esse fenômeno não é tão expressivo quanto o *chaining*, mas outros exemplos do mesmo podem ser encontrados nos casos testados, presentes no anexo.

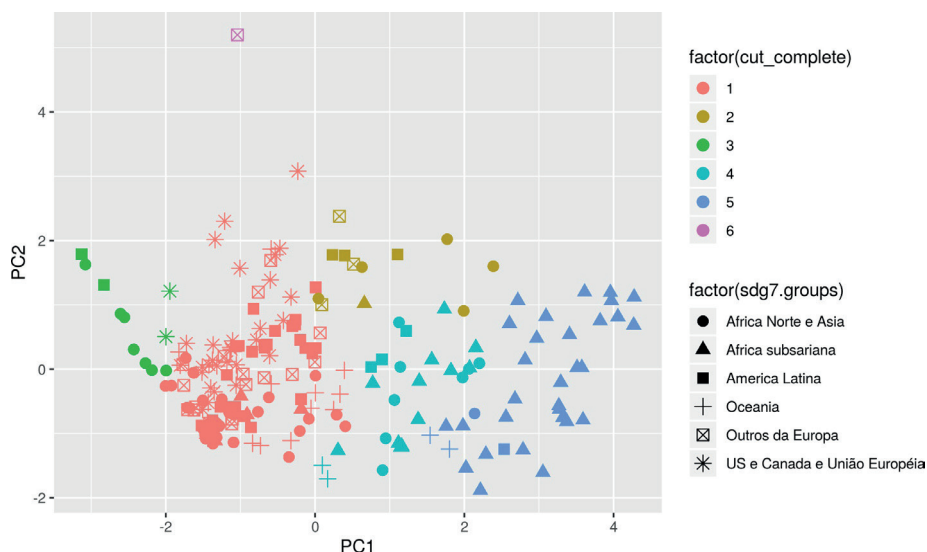



Figura 3.11: Agrupamento com o método *Average*.

3.6 MÉTODO HIERÁRQUICO DIVISIVO

Dentre as metodologias hierárquicas, uma delas é a divisiva. Nesta abordagem, são feitas divisões sucessivas das amostras (segundo alguma métrica) a fim de se obter mais e mais divisões nos grupos. Para testar essa abordagem utilizados o pacote uclust (Cybis et al., 2018), do R, que divide tantas vezes quanto for estatisticamente relevante. Como podemos ver nas Figuras 3.12 e 3.13, essa abordagem gera muitos grupos.

Usando a métrica euclidiana temos um agrupamento com 39 grupos, e usando a métrica euclidiana ao quadrado (padrão do pacote) temos 35 grupos. Consideramos essa quantidade demasiado grande para ser interpretada.

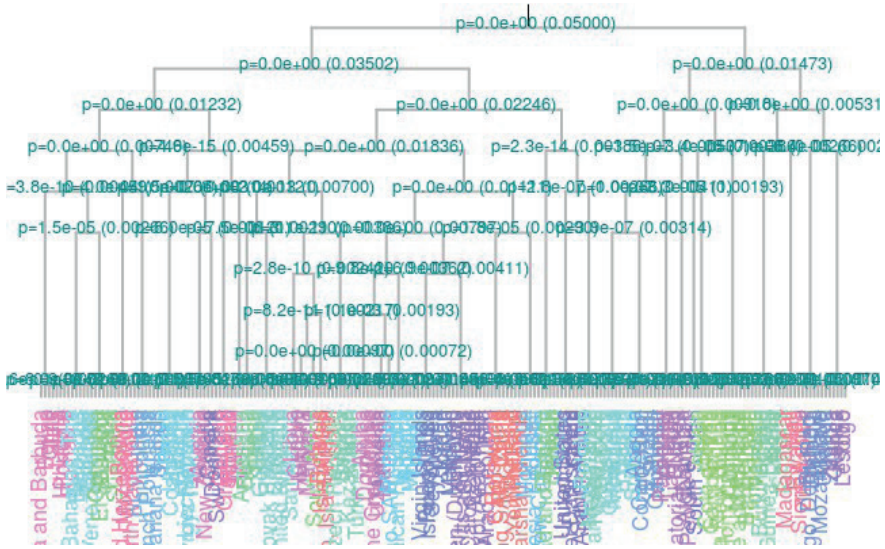



Figura 3.12: Agrupamento com o método U-clust, usando a métrica euclidiana.

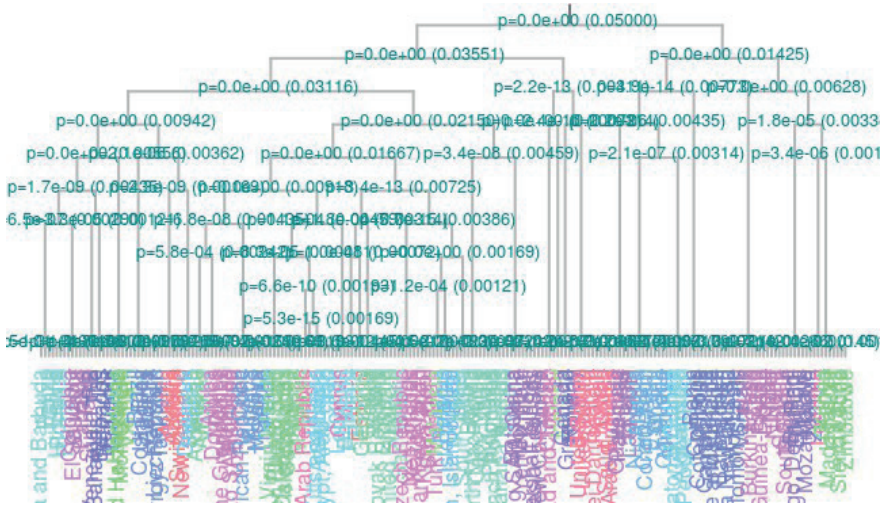



Figura 3.13: Agrupamento com o método U-clust, usando a métrica euclidiana ao quadrado, default do pacote.

3.7 KNN APLICADO PARA O IDH

O índice de Desenvolvimento Humano (IDH) é um índice estatístico composto que busca medir a qualidade de vida, estabelecendo um parâmetro comparativo entre países ou regiões a partir de indicadores de renda nacional bruta (RNB) per capita, anos de escolaridade e expectativa de vida. O índice foi proposto em 1990 pelo economista Amartya Sen e, anualmente, a Organização das Nações Unidas (ONU) elabora uma lista classificando seus países membros segundo o IDH. Apesar de sofrer algumas críticas, como a falta de consideração com indicadores ambientais, o IDH é tido como a melhor referência disponível atualmente para retratar a qualidade de vida de uma população estudada. Para o presente trabalho, os países foram categorizados com desenvolvimento humano elevado (IDH acima de 0,872), baixo (IDH abaixo de 0,550) e médio (IDH entre 0,872 e 0,550), conforme o último relatório da ONU (2019).

Na literatura econômica, encontra-se frequentemente indícios de relação de crescimento econômico e oferta e consumo energético (Payne, 2010), associados à promoção do desenvolvimento econômico e social - seguindo o trabalho pioneiro de Kraft and Kraft (1978) - pelas facilidades que proporciona, como a redução do esforço físico nas atividades produtivas, além de outras possibilidades relacionadas à segurança alimentar, saúde, educação e lazer. Mais recentemente, sendo a agenda de política de desenvolvimento de energia renovável uma recomendação da ONU (2007) com o intuito de amparar o desenvolvimento social, econômico e ambiental de longo prazo no mundo, vários estudos (Bhattacharya et al., 2016) almejaram verificar a relação entre energia renovável e crescimento econômico.

O método a ser aplicado nessa seção é o KNN ou "*k*-Nearest Neighbor" (*k* vizinho mais próximo). Esse método não-paramétrico pode ser usado tanto para classificação como regressão. O centro de seu funcionamento está em um esquema de votação para uma dada amostra, atribuindo a ela pela classe dos *k* vizinhos mais próximos pela distância euclidiana ou de Manhattan. Dentre as vantagens frequentemente apontadas pelo método são sua simplicidade e fácil implementação e inexistência de premissas, enquanto apresenta pontos negativos como sensibilidade a outliers, falta de capacidade em lidar com missing data e seu melhor funcionamento com poucas dimensões.

Primeiramente, far-se-á um KNN de modo a verificar se os seis grupos encontrados pelo *k*-means seriam representados também pelos seis vizinhos próximos, pois nesse patamar de *k* já se obtém boa acurácia e certa estabilidade. Inicialmente, foram consideradas todos os indicadores obteve-se que para 94,71% dos países chegou-se à mesma classificação do método que prioriza a média. Almejando facilitar a visualização e melhorar o desempenho do KNN, realizou-se novamente o procedimento agora com PCA1 e PCA2, que obteve desempenho de 92,31%, levemente inferior quando comparado ao de maior dimensão, porém, mais simples. Na Figura 3.14, estão representadas as seis regiões de classificação pelo KNN, enquanto as cores dos pontos referem-se ao grupo extraído do *k*-means.

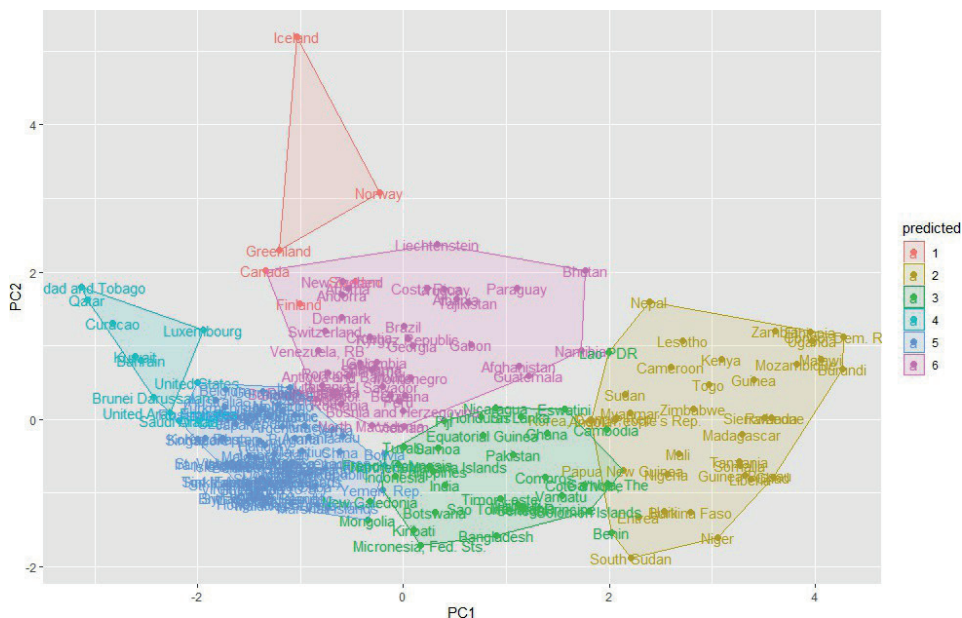



Figura 3.14: A KNN para grupos do k-means = 6.

De modo posterior, foi realizado em um primeiro momento o KNN utilizando todos os indicadores disponíveis para tentar realizar uma classificação condizente com as três faixas do IDH dos países, seguindo o procedimento anterior. Um resumo dos indicadores, por IDH, pode ser encontrado na Tabela 3.8. Nessa tentativa, a acurácia foi de 85,58% e o modelo teve dificuldade em diferenciar principalmente países na África Subsariana de médio e baixo IDH, que apresentam percentual de fonte energética renovável e de uso de energia em níveis similares, mas diferenças notáveis em acesso à eletricidade e a combustíveis limpos.

Tabela 3.8: Indicadores dos países por IDH.

Países	DH		ACLTC	AE	NIEEP	CER	PER	UE
31	Alto	Média	95,36	98,05	4,48	21,95	39,82	4.803,27
		Desvio	17,38	10,85	2,69	19,25	31,75	2.928,05
142	Médio	Média	72,74	91,56	5,21	23,75	31,46	1.965,87
		Desvio	30,12	16,18	2,75	23,24	28,76	2.515,76
35	Baixo	Média	15,37	37,75	7,79	63,12	42,77	678,16
		Desvio	24,41	22,42	5,45	27,67	34,50	446,40

Nota: ACLTC = Acesso a combustíveis limpos e tecnologias para cozinhar; AE = Acesso a eletricidade; NIEEP = Nível de intensidade energética da energia primária; CER = Consumo de energia renovável; PER = Produção de eletricidade renovável; UE = Uso de energia

Em seguida, reduziu-se a dimensionalidade pelas duas componentes principais. Na Figura 3.15, estao representadas as três regiões de classificação pelo KNN, enquanto as cores dos pontos referem-se ao grupo extraído do *k*-means. Nesse caso, o KNN classificou corretamente o IDH em 80,77%, sendo que o modelo apresentou erro em diferenciar 15 dos 31 dos países que possuem alto desenvolvimento humano, classificando-os como médio, afetado pelo desvio padrão de 31,75% na produção de energia renovável na matriz energética desses países, variável de maior relevância no PCA2. Em particular, enquanto Liechtenstein, Noruega e Islândia apresentam mais de 95% de sua energia oriunda de fontes limpas, Israel, Hong Kong e Cingapura apresentam menos de 2% no mesmo indicador, possuindo mais vizinhos de médio desenvolvimento, principalmente países da Ásia e América Latina. Por outro lado, os países com IDH mais elevado compartilham a menor variabilidade nos indicadores de acesso à eletricidade e a combustíveis limpos. A classificação do KNN ainda permite verificar o grande grupo de países de desenvolvimento intermediário, que abrange países com grande desigualdade de acesso a combustíveis limpos e à eletricidade, sendo observado um contraste entre países europeus, com alto acesso, contra países africanos e da Oceania.

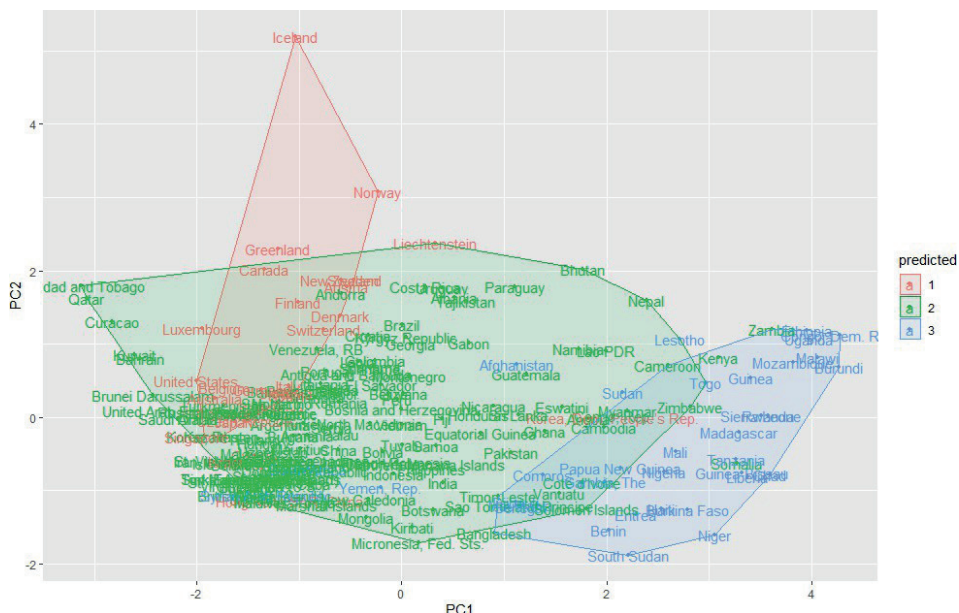



Figura 3.15: Classificação de IDH por KNN.

CONCLUSÃO

O presente trabalho utilizou diferentes métodos de agrupamento visando agrupar países através da utilização dos indicadores de desenvolvimento sustentável (SDG's) propostos pela ONU em 2015. Neste trabalho, o foco foi no SDG 7 - Energia limpa e acessível.

Os métodos empregados foram PCA para redução de dimensionalidade e *k*-means e modelos hierárquicos agrupamento. Diferentes métricas de análise foram propostas para avaliação do método. O modelo de aprendizagem para o agrupamento é não supervisionado, ou seja, não há uma resposta correta para que o modelo aprenda, de modo que os resultados dependem da visão do pesquisador para análise e conclusões relevantes ao campo de estudo. Ainda, para complementar a análise, o método de KNN para classificação, foi usado para classificar de acordo com os grupos encontrados no trabalho. De todos os métodos empregados, o *k*-means se destacou por apresentar grupos coerentes com o cenário global e por este motivo apresenta uma análise mais detalhada.

O agrupamento de países a partir de indicadores de desenvolvimento sustentável possibilitam uma análise que foge aos padrões de agrupamento usuais, como agrupamentos por localização geográfica através dos continentes ou por PIB. A separação a partir de dois principais indicadores, um que observa o desenvolvimento energético do país e o outro voltado para a utilização de fontes renováveis, permitiu a visualização de características comuns comprovadas pela análise das matrizes energéticas por fonte. Na análise das matrizes por fonte não apenas a presença renovável é importante, mas também a capacidade de geração daquele país, ligada ao seu desenvolvimento energético.

ANEXOS

No que segue são apresentados os gráficos e tabelas que resumem os resultados para as diferentes clusterizações consideradas neste trabalho.

Agrupamento com métrica Euclidiana

O método de agrupamento *ward.D* e *ward.D2* resultaram em mais divisões, com países com geopolítica aparentemente próximos.

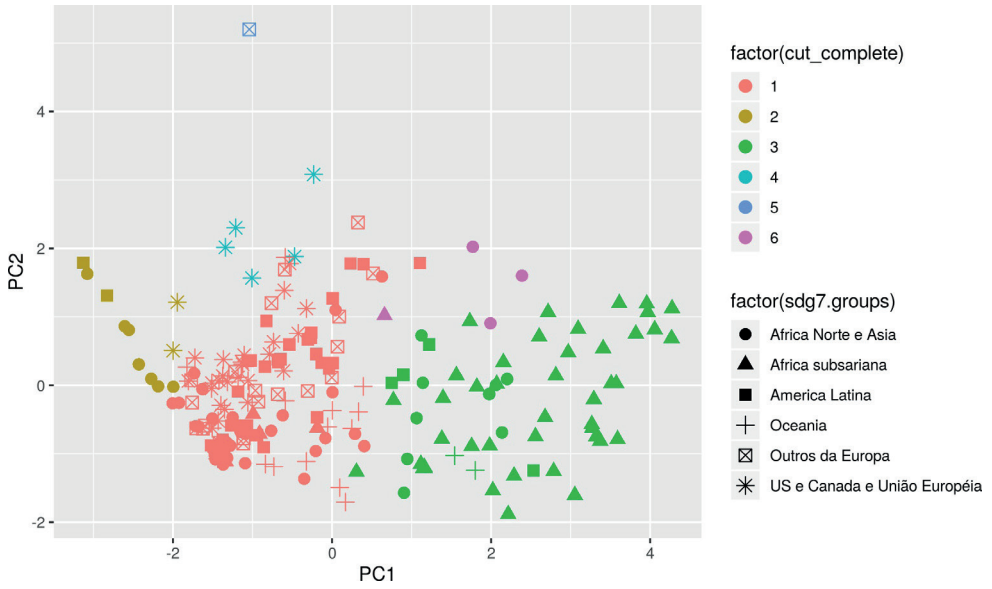



Figura 3.16: Agrupamento com o método *Average*.

Figura 3.17: Agrupamento com o método *Centroid*.

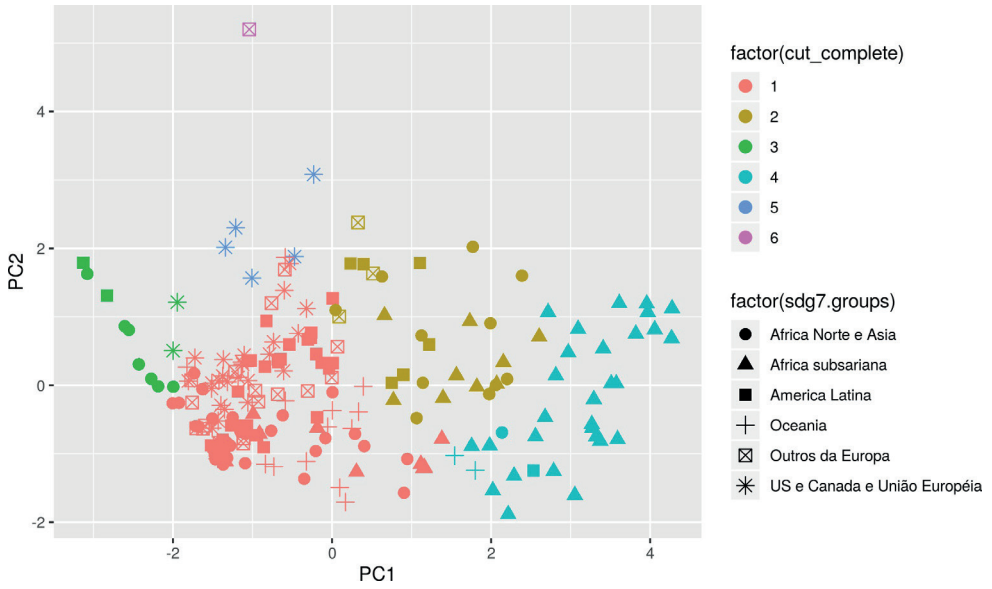



Figura 3.18: Agrupamento com o método *Complete*.

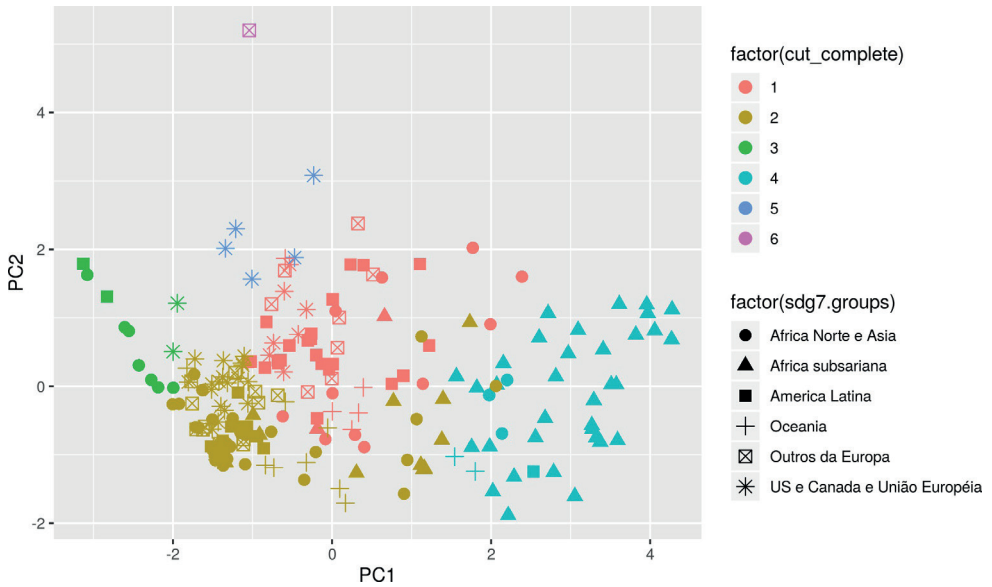



Figura 3.19: Agrupamento com o método *Mcquitty*.

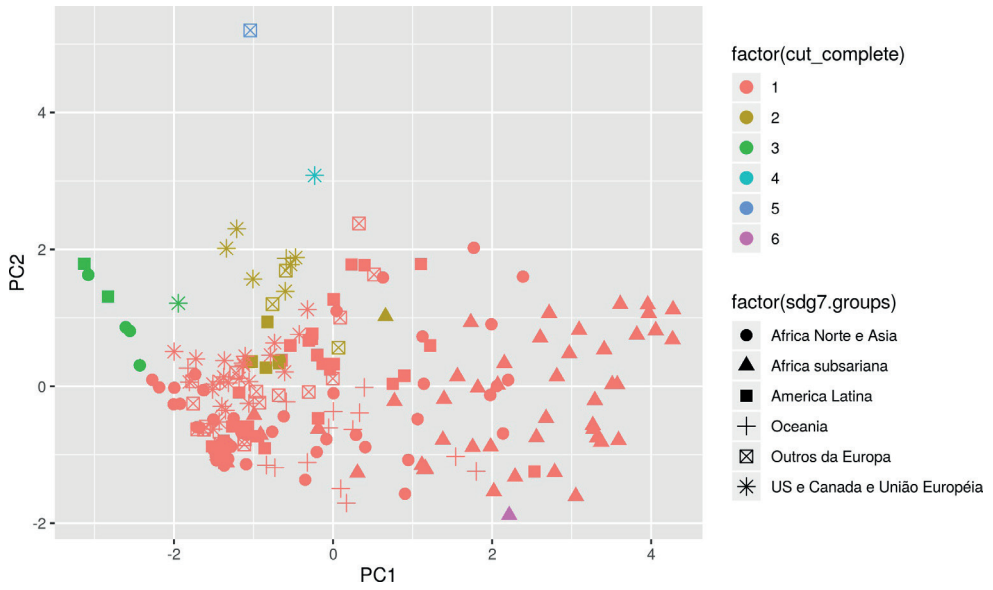



Figura 3.20: Agrupamento com o método *Median*.

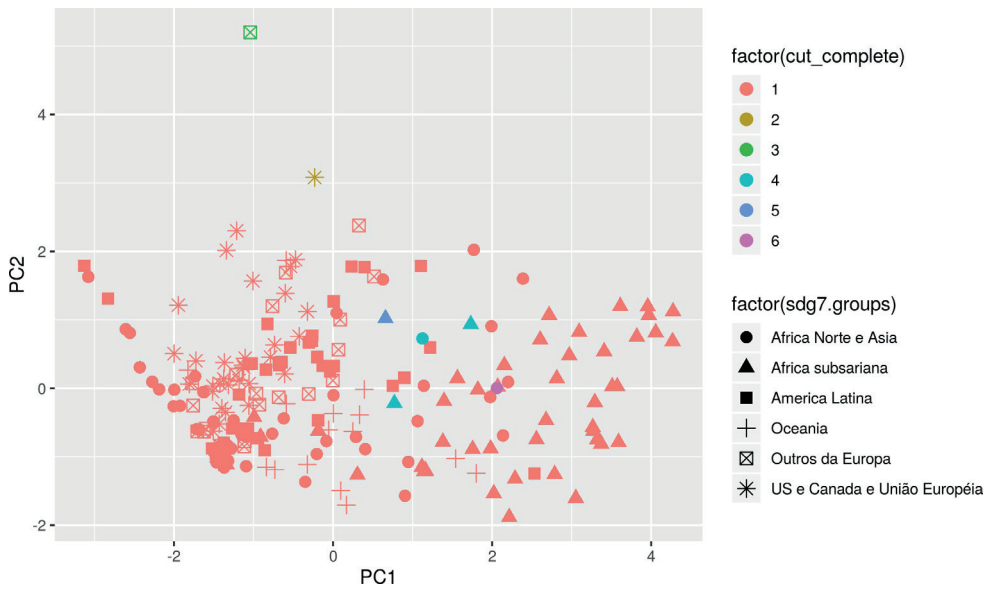



Figura 3.21: Agrupamento com o método *Single*.

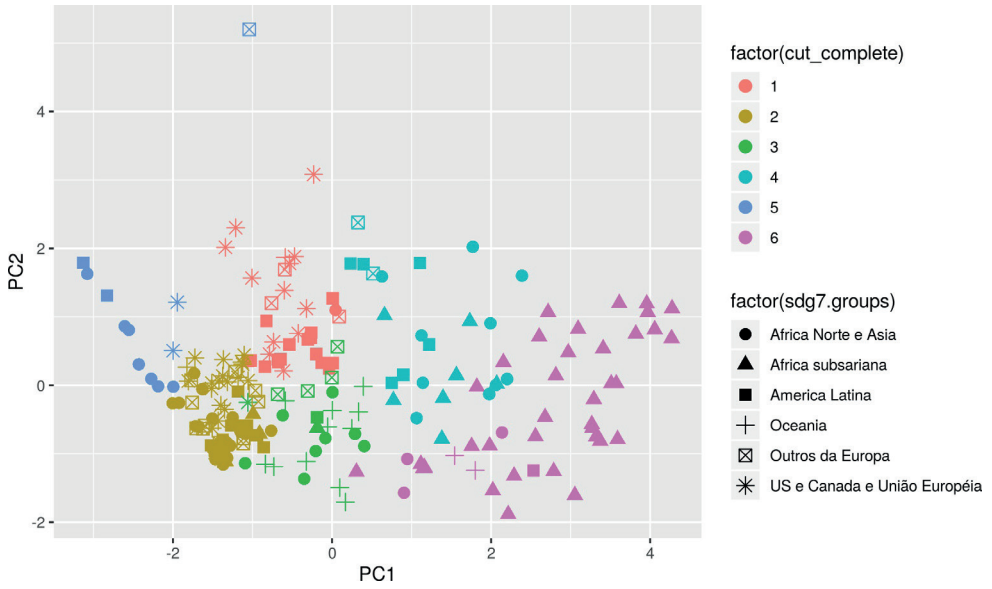



Figura 3.22: Agrupamento com o método *Ward.D*.

Agrupamento utilizando a métrica Manhattan

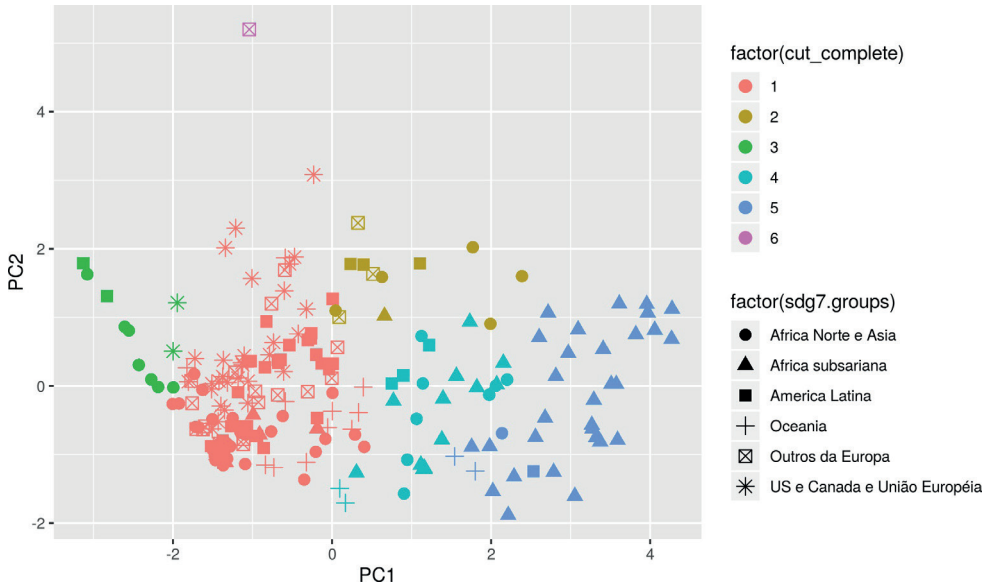



Figura 3.23: Agrupamento com o método *Average*.

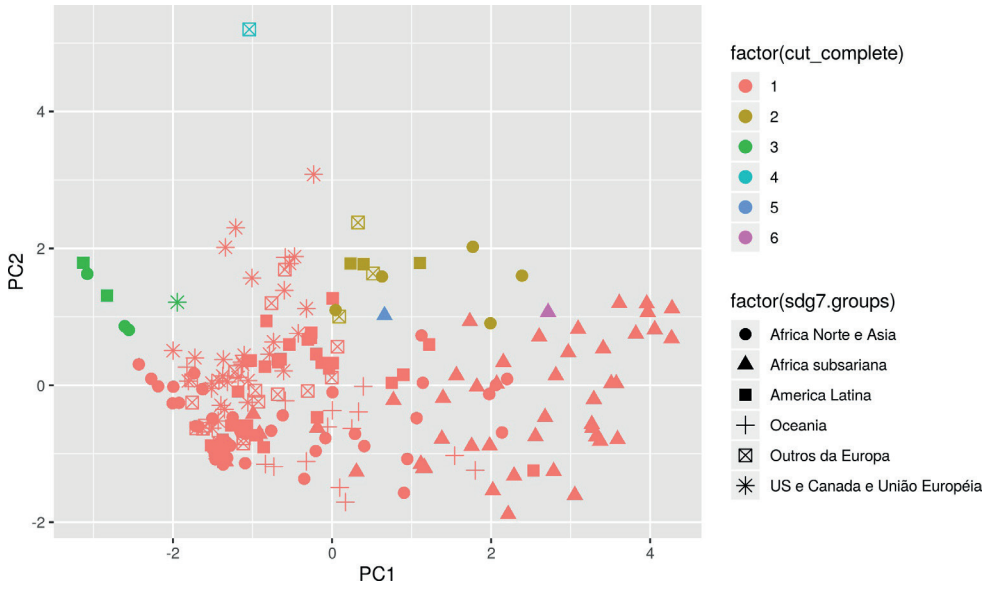



Figura 3.24: Agrupamento com o método *Centroid*.

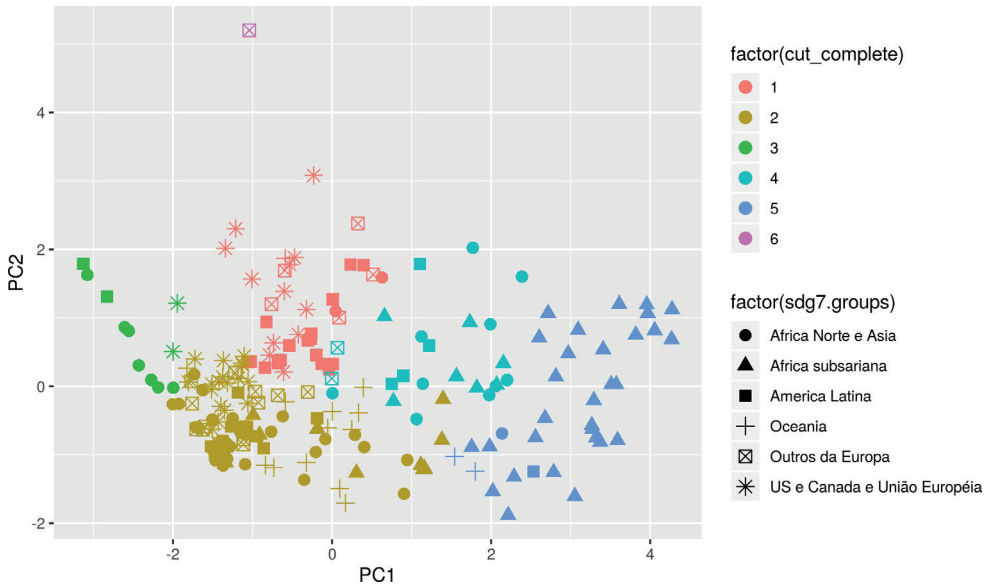



Figura 3.25: Agrupamento com o método *Complete*.

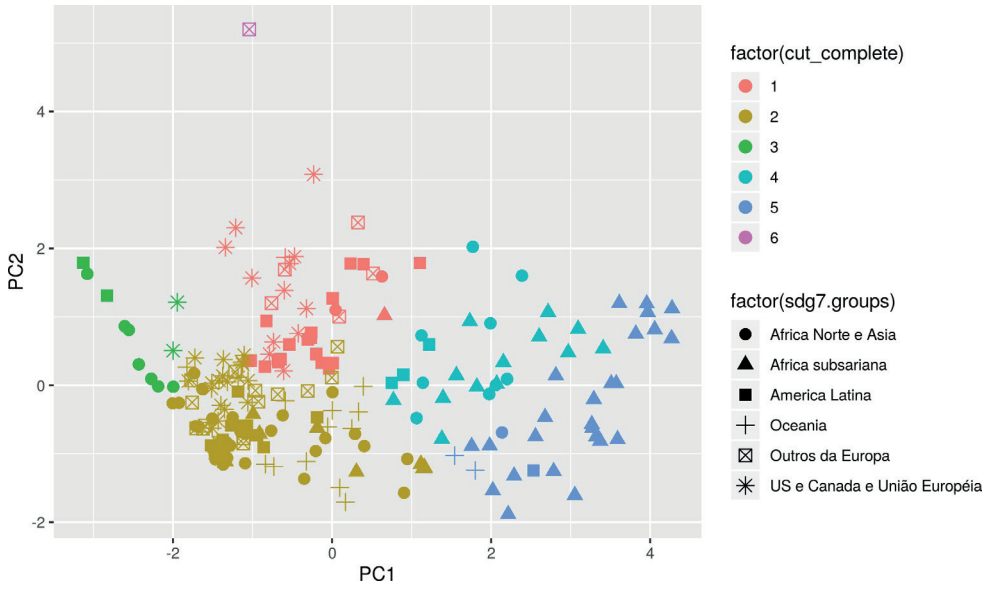



Figura 3.26: Agrupamento com o método *Mcquitty*.

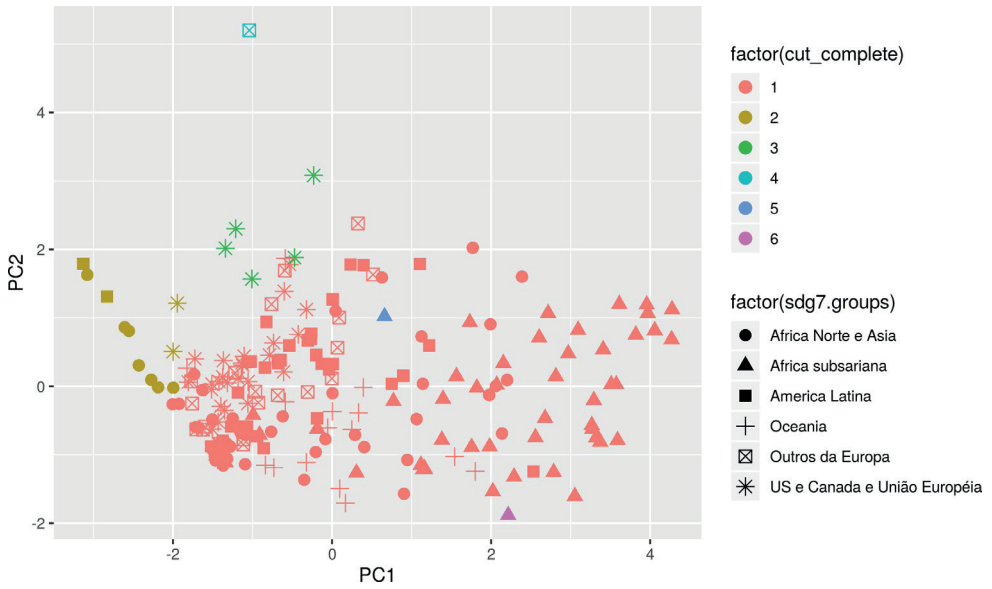



Figura 3.27: Agrupamento com o método *Median*.

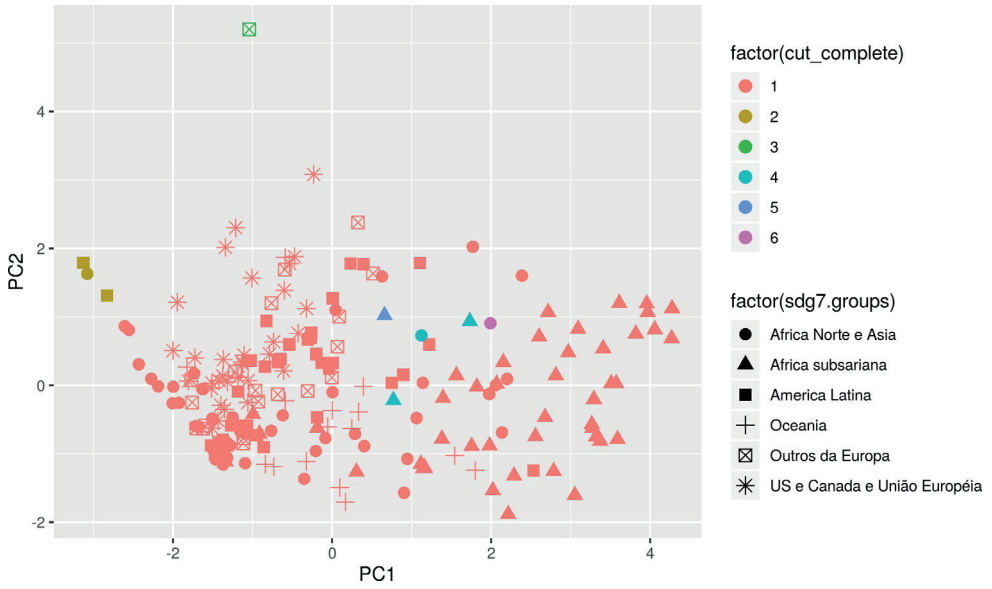



Figura 3.28: Agrupamento com o método *Single*.

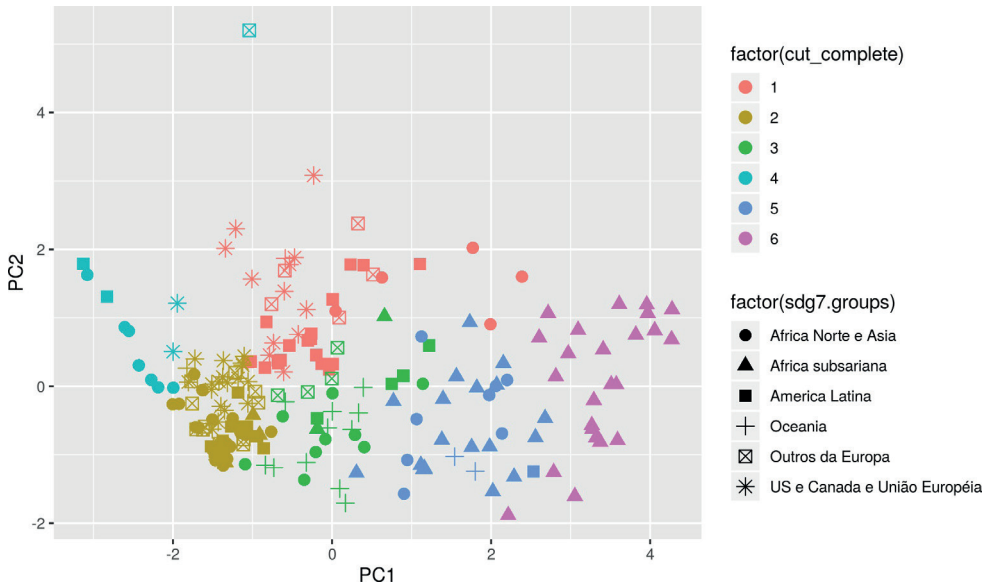



Figura 3.29: Agrupamento com o método *Ward.D*.

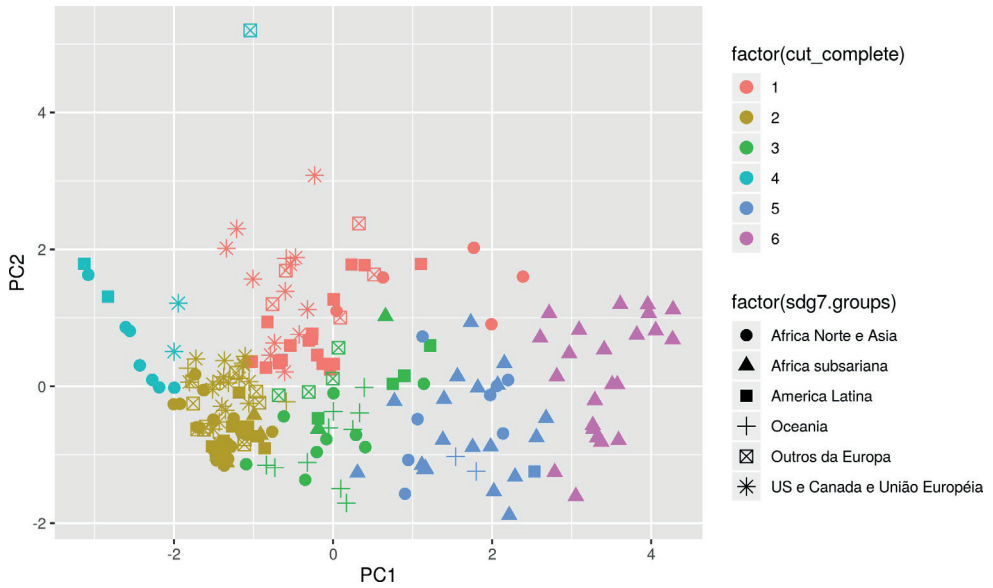



Figura 3.30: Agrupamento com o método *Ward.D2*.

Agrupamento com métrica Maximum

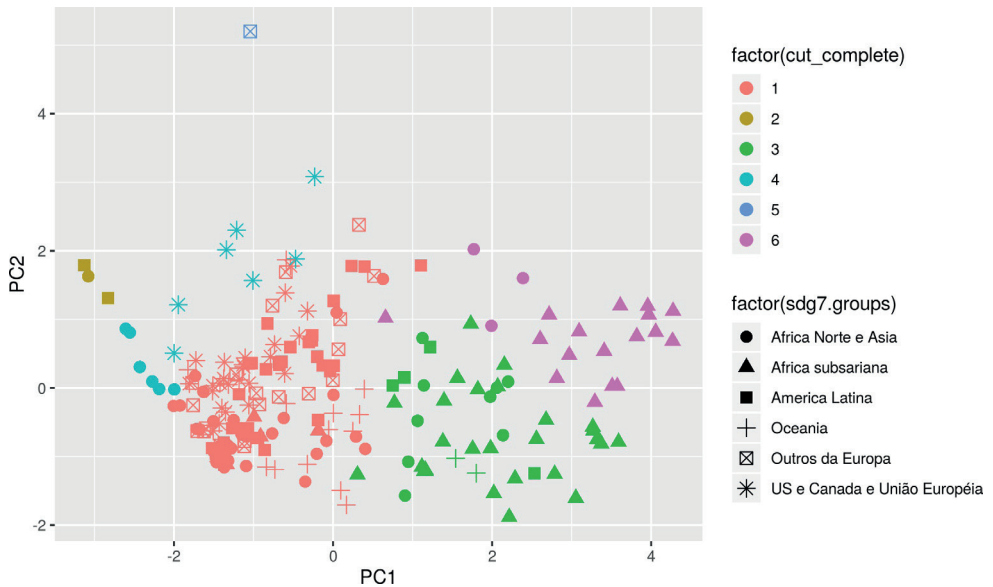



Figura 3.31: Agrupamento com o método *Average*.

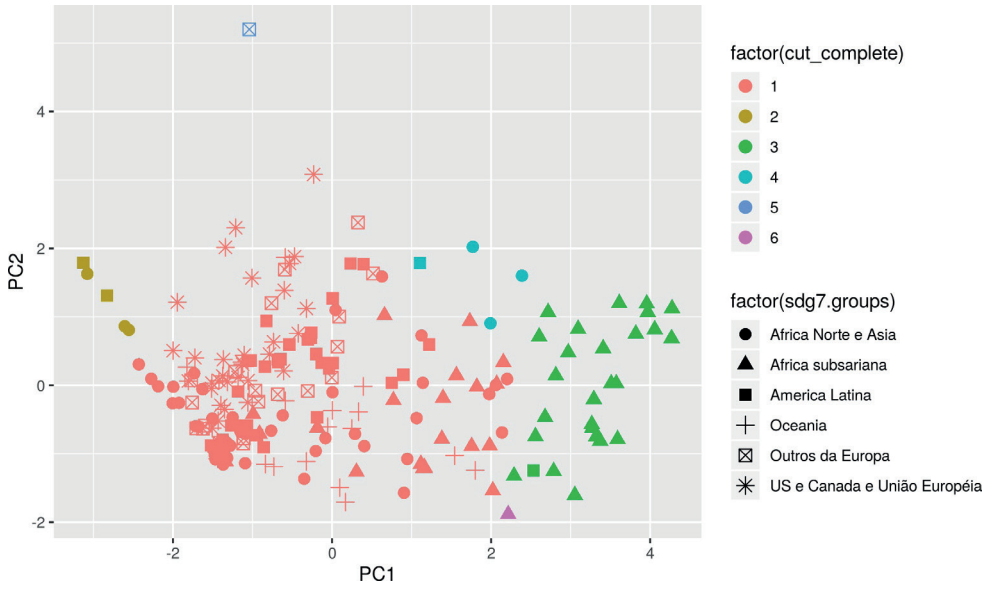



Figura 3.32: Agrupamento com o método *Centroid*.

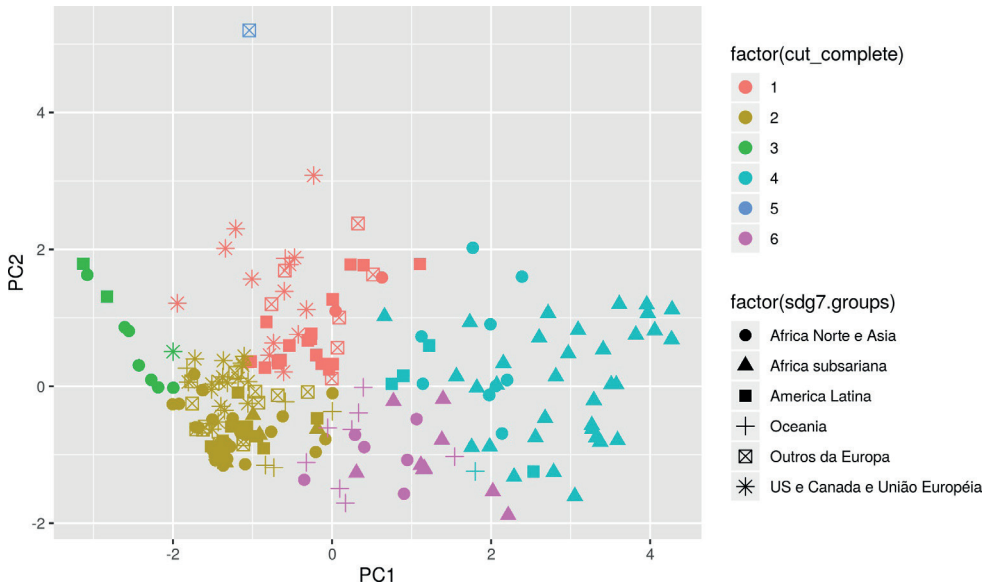



Figura 3.33: Agrupamento com o método *Complete*.

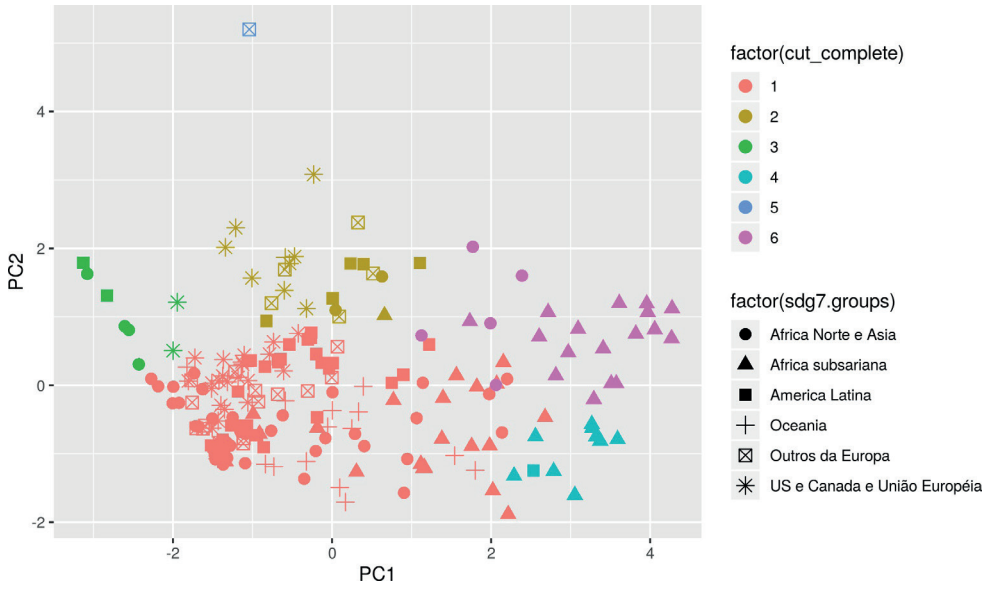



Figura 3.34: Agrupamento com o método *Mcquitty*.

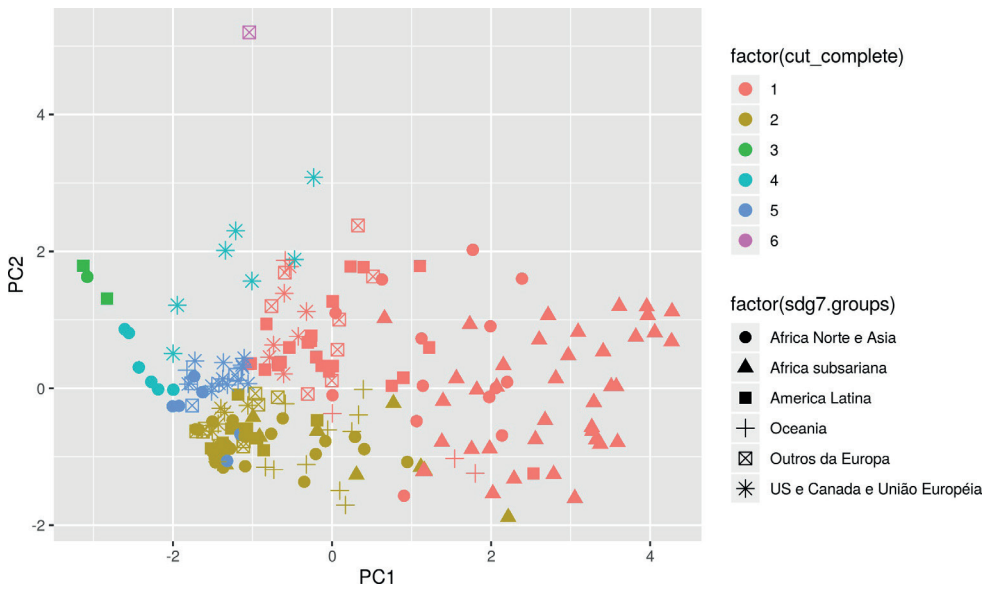



Figura 3.35: Agrupamento com o método *Median*.

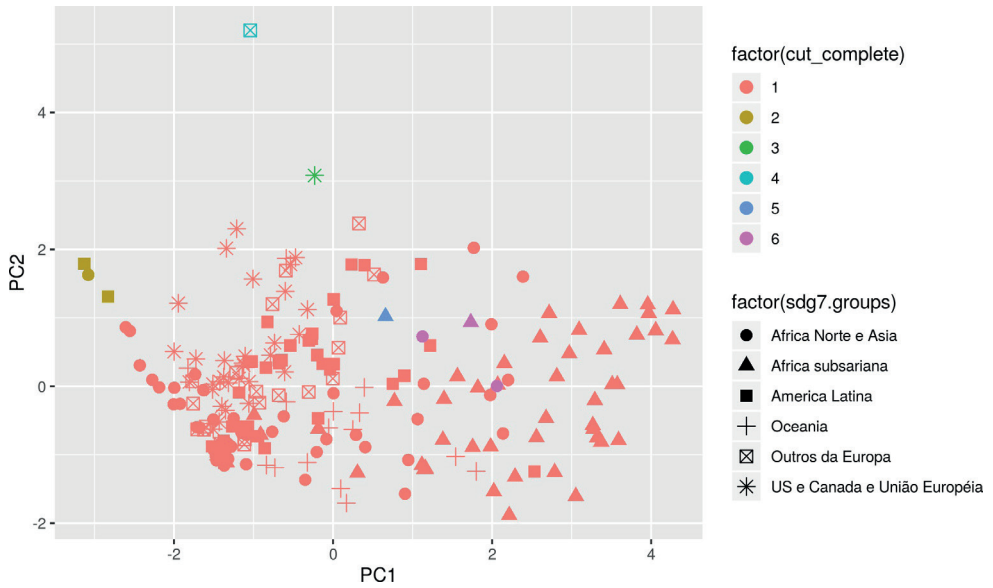



Figura 3.36: Agrupamento com o método *Single*.

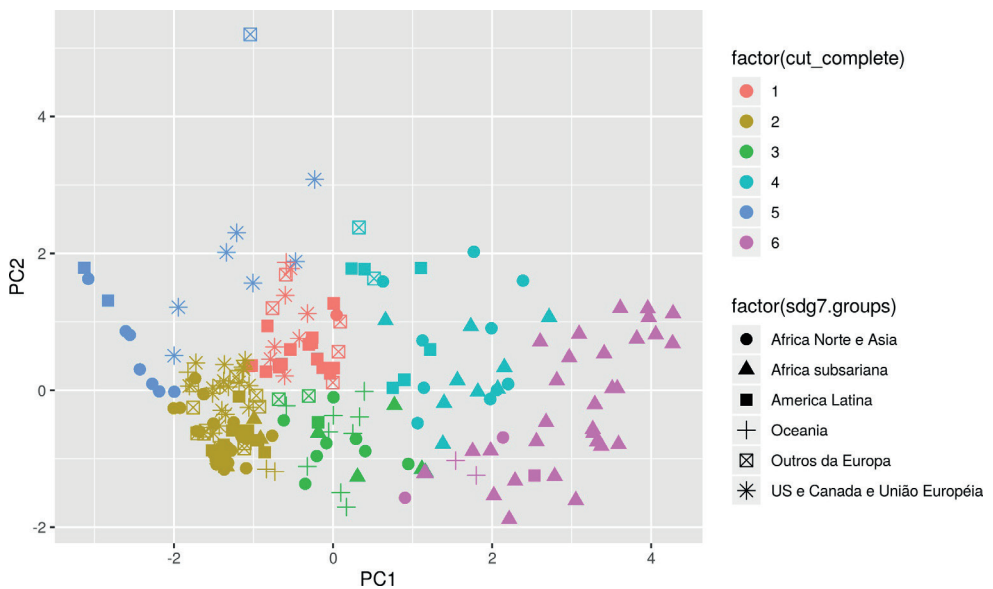



Figura 3.37: Agrupamento com o método *Ward.D*.

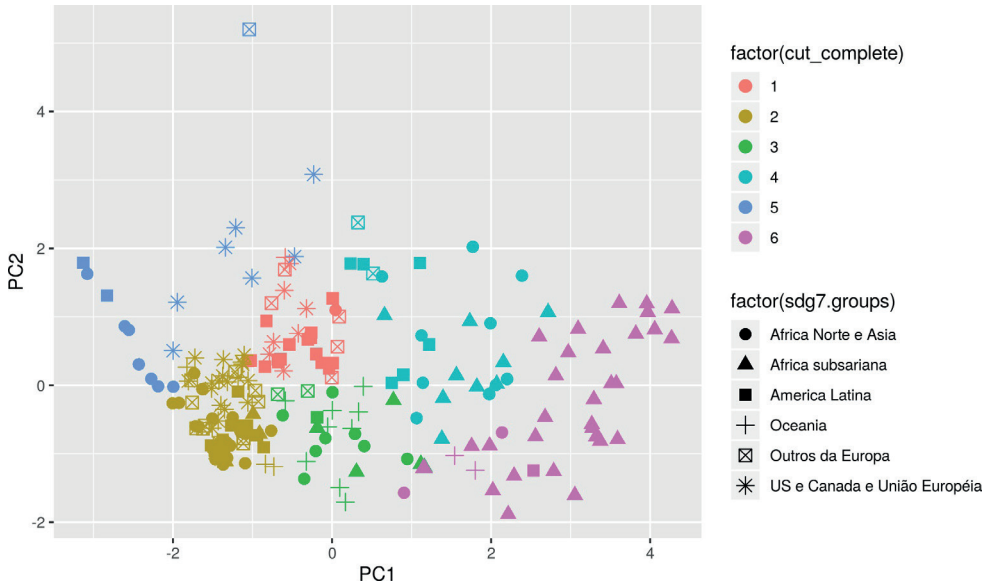



Figura 3.38: Agrupamento com o método *Ward.D2*.

Agrupamento com métrica Minkowsky

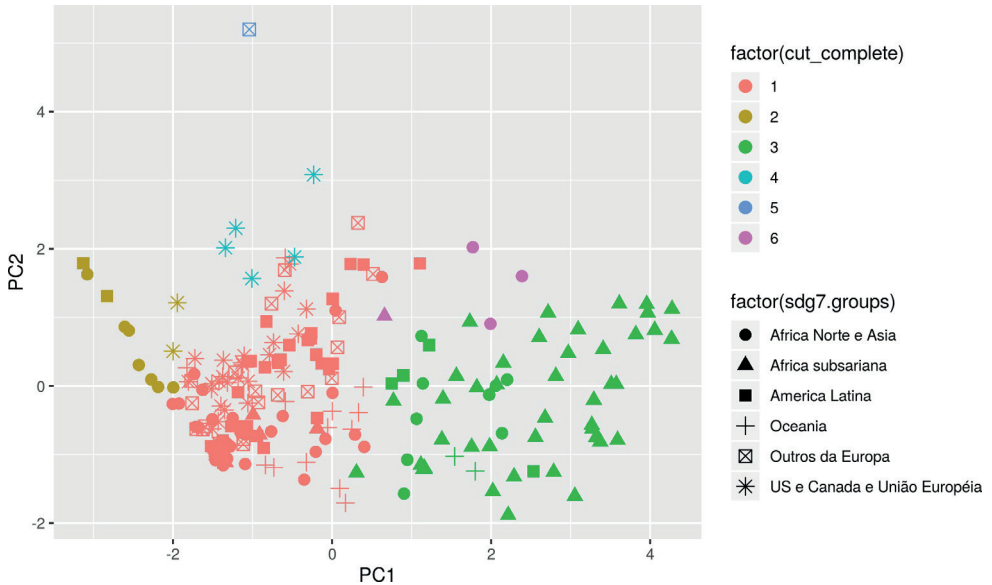



Figura 3.39: Agrupamento com o método *Average*.

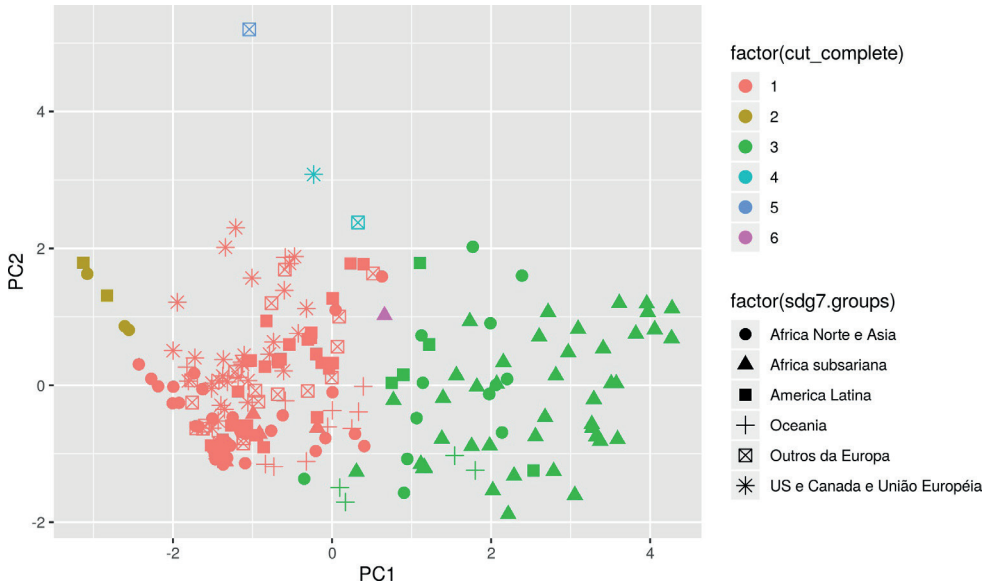



Figura 3.40: Agrupamento com o método *Centroid*.

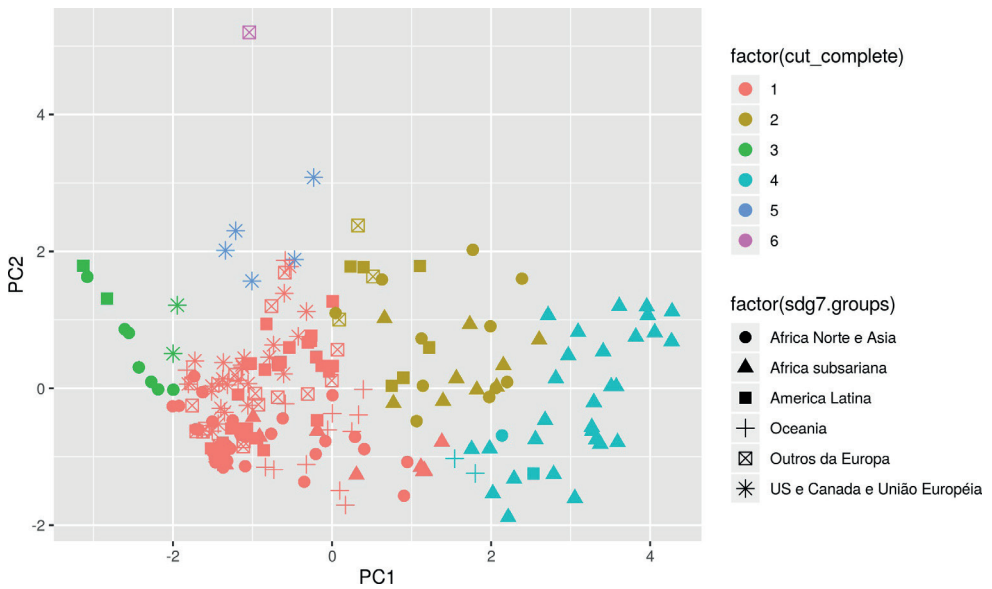



Figura 3.41: Agrupamento com o método *Complete*.

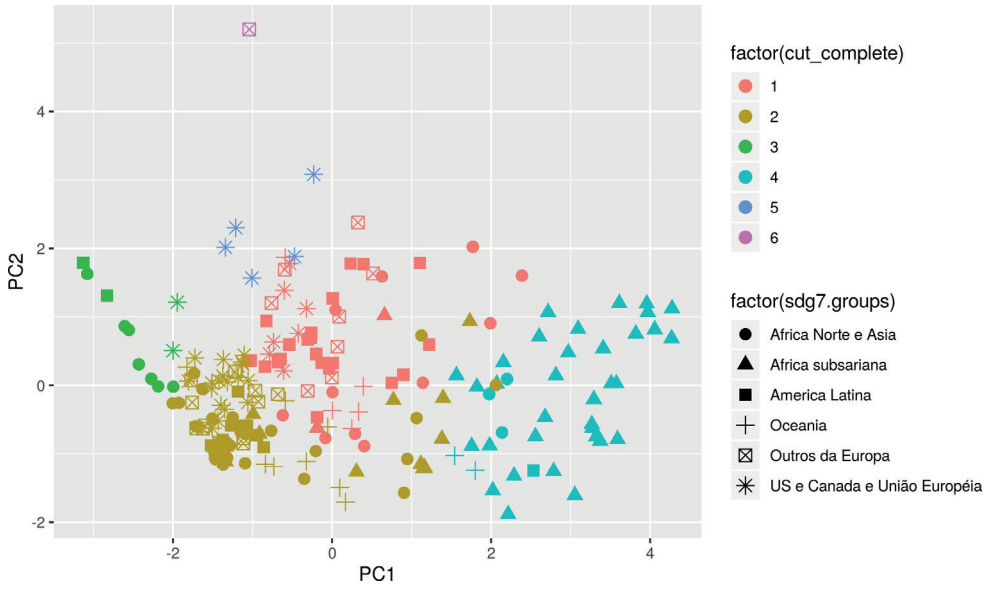



Figura 3.42: Agrupamento com o método *Mcquitty*.

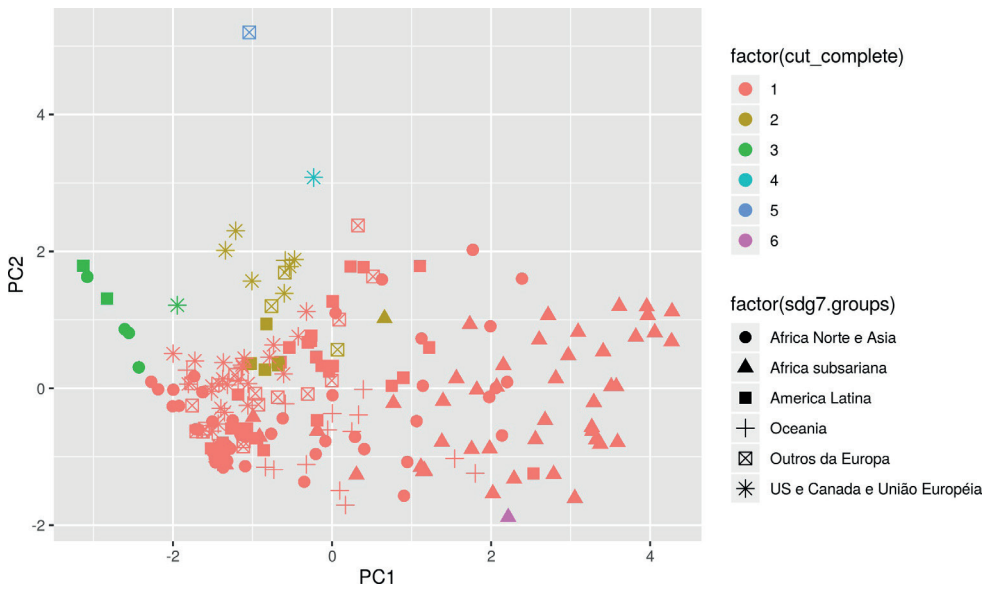



Figura 3.43: Agrupamento com o método *Median*.

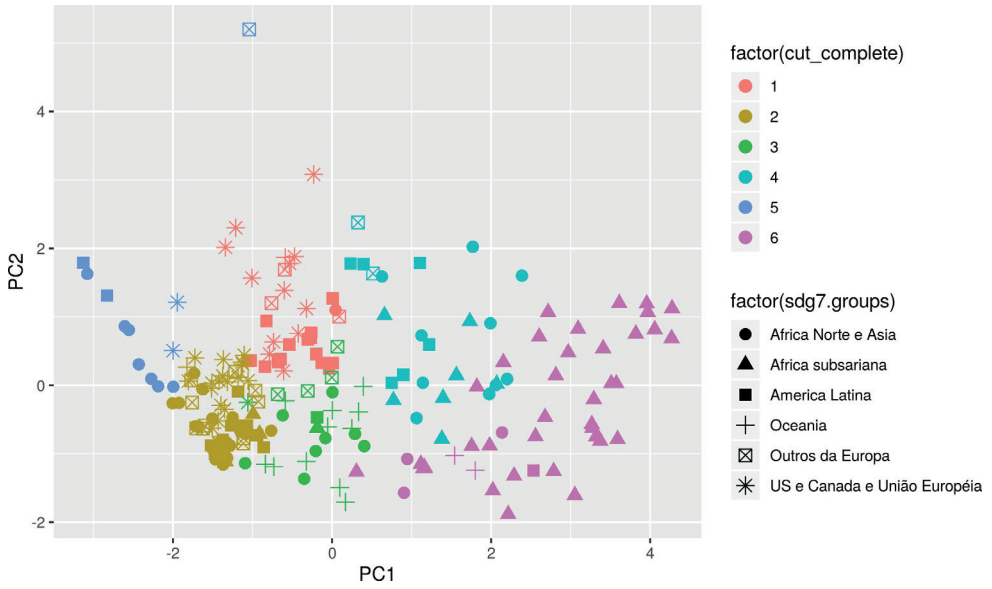



Figura 3.44: Agrupamento com o método *Ward.D*

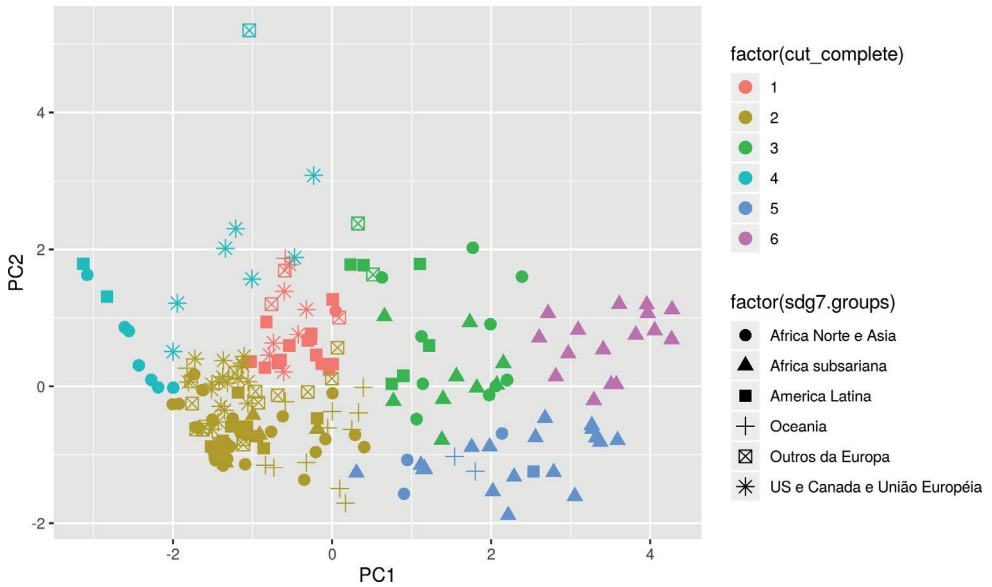



Figura 3.45: Agrupamento com o método *Ward.D2*

Resultados para os agrupamentos *k-means*

No cabeçalho das tabelas que seguem adota-se as seguintes abreviações

ACLTC: Acesso a combustíveis limpos e tecnologias para cozinhar

AE: Acesso a eletricidade

NIEEP: Nível de intensidade de energia da energia primária

CER: Consumo de energia renovável

PER: Produção de eletricidade renovável

UE: Uso de energia

Tabela 3.9: Agrupamento *k*-means - Dados para o Grupo 1

	ACLTC	AE	NIEEP	CER	PER	UE
Canada	100	100	7,34	22,03	63,01	7.788,6
Greenland	100	100	6,37	15,53	81,29	7.161,5
Finland	100	100	6,37	43,24	44,50	6.828,8
Norway	100	100	3,75	57,77	97,71	6.934,6
Sweden	100	100	4,27	53,25	63,26	5.427,9
Iceland	100	100	16,56	77,03	99,98	17.023,2

Tabela 3.10: Agrupamento *k*-means - Dados para o Grupo 2

	ACLTC	AE	NIEEP	CER	PER	UE
Haiti	4,27	40,91	10,11	76,07	8,00	381,59
Angola	47,36	42,00	3,61	49,57	53,17	520,96
Burkina Faso	8,46	21,79	6,03	74,17	9,35	685,18
Burundi	0,83	8,47	7,72	95,68	82,71	685,18
Cameroon	21,98	58,81	4,78	76,54	76,12	342,50
Chad	3,13	7,70	2,78	89,36	26,64	685,18
Congo, Dem. Rep.	3,94	16,83	20,94	95,82	99,82	307,42
Congo, Rep,	22,74	60,40	4,02	62,40	53,34	392,37
Eritrea	15,60	45,88	4,83	79,77	0,49	233,66
Ethiopia	3,40	29,00	13,67	92,16	99,96	486,69
Guinea	1,24	31,62	10,62	76,27	78,75	685,18
Guinea-Bissau	1,46	20,30	12,02	86,85	26,64	685,18
Kenya	12,76	41,60	7,85	72,66	87,51	464,41
Lesotho	34,74	29,97	9,72	52,14	100,00	685,18
Liberia	0,70	15,34	25,99	83,85	26,64	685,18
Madagascar	0,90	20,01	5,37	70,17	54,60	685,18
Malawi	2,49	10,80	4,08	83,65	91,31	685,18

Mali	0,93	37,60	2,83	61,53	43,52	685,18
Mozambique	3,62	24,00	17,31	86,40	86,41	423,27
Niger	1,88	16,60	6,95	78,94	0,75	135,36
Nigeria	4,39	52,50	5,68	86,64	18,20	756,35
Rwanda	0,56	22,80	4,88	86,66	56,89	685,18
Sierra Leone	0,93	19,18	7,00	77,66	60,98	685,18
Somalia	2,12	29,39	5,01	94,29	26,64	685,18
South Sudan	0,63	18,60	1,09	39,07	0,61	685,18
Sudan	38,95	49,26	3,96	61,60	64,54	361,89
Tanzania	2,15	26,57	8,34	85,71	34,15	465,93
Togo	6,24	45,19	14,34	71,26	75,31	485,11
Uganda	0,76	18,50	9,64	89,06	92,95	1.357,74
Zambia	16,31	31,10	7,34	87,99	96,99	604,41
Zimbabwe	29,36	33,70	15,80	81,80	52,72	744,93
Korea, Dem. People's Rep.	10,08	39,57	6,07	23,12	72,80	770,13
Myanmar	16,97	60,50	3,12	61,53	58,85	277,09
Nepal	26,56	86,66	7,42	85,26	100,00	377,99
Papua New Guinea	12,85	44,55	9,28	52,50	34,53	515,20

Tabela 3.11: Agrupamento *k*-means - Dados para o Grupo 3

	ACLTC	AE	NIEEP	CER	PER	UE
Honduras	51,87	89,98	6,15	51,54	42,28	548,73
Nicaragua	51,41	83,63	5,43	48,20	50,05	507,57
Benin	6,06	40,16	9,07	50,86	5,56	397,14
Botswana	63,15	58,35	3,35	28,88	0,03	1.083,91
Comoros	8,15	74,91	4,69	45,33	26,64	685,18
Cote d'Ivoire	17,96	63,31	7,24	64,53	16,73	494,95
Djibouti	10,89	57,55	3,41	15,38	26,64	685,18
Equatorial Guinea	34,00	66,59	2,21	7,82	57,83	685,18
Eswatini	48,52	66,03	4,61	66,10	46,57	685,18
Gambia, The	3,23	53,63	4,50	51,51	26,64	685,18
Ghana	20,37	75,85	3,75	41,41	50,89	298,91
Mauritania	45,66	39,50	3,59	32,16	13,37	685,18
Sao Tome and Principe	17,88	67,60	4,66	41,06	10,47	685,18
Senegal	31,82	60,50	4,98	42,71	10,42	302,19
Bangladesh	16,68	73,17	3,14	34,75	1,23	206,76
Cambodia	16,36	69,03	5,77	64,92	46,42	370,26
India	39,88	88,00	4,73	36,02	15,34	561,65
Indonesia	56,49	97,54	3,53	36,88	10,65	877,08

Lao PDR	5,54	89,70	5,17	59,32	86,37	515,20
Mongolia	41,32	82,74	6,10	3,43	3,08	1.449,47
Pakistan	42,13	71,20	4,42	46,48	31,43	473,85
Philippines	43,21	89,08	3,12	27,45	25,41	429,91
Sri Lanka	25,44	93,97	2,06	52,88	48,48	480,77
Timor-Leste	6,53	67,28	6,07	18,22	23,39	1.234,30
Fiji	39,16	95,22	4,85	31,26	45,02	1.234,30
French Polynesia	26,86	100,00	6,07	9,83	32,01	1.234,30
Kiribati	4,90	90,56	4,14	4,25	7,27	1.234,30
Micronesia, Fed. Sts.	11,79	76,11	6,59	1,20	1,60	.1234,30
New Caledonia	26,86	100,00	5,01	4,76	14,09	1.234,30
Northern Mariana Islands	26,86	100,00	5,01	30,41	23,39	1.234,30
Samoa	31,53	99,50	5,21	34,32	30,35	1.234,30
Solomon Islands	8,36	55,10	5,02	63,31	2,26	1.234,30
Tuvalu	48,50	99,17	3,91	30,41	28,17	1.234,30
Vanuatu	12,72	48,42	3,87	36,11	21,26	1.234,30

Tabela 3.12: Agrupamento *k*-means - Dados para o Grupo 4

	ACLTC	AE	NIEEP	CER	PER	UE
Curacao	86,35	100,00	3,87	0,35	3,71	13.710,08
Trinidad and Tobago	99,27	100,00	19,09	0,28	8,91	15.108,65
United States	100,00	100,00	5,41	8,72	13,23	7.161,45
Luxembourg	100,00	100,00	2,87	9,03	32,38	8.329,48
Bahrain	100,00	100,00	9,79	2,53	5,09	10.207,87
Brunei Darussalam	100,00	100,00	3,65	0,01	0,05	8.337,49
Kuwait	100,00	100,00	5,32	4,09	2,92	10.721,49
Oman	95,06	100,00	6,30	4,09	2,92	6.154,48
Qatar	98,44	100,00	6,40	4,09	2,92	14.890,44
Saudi Arabia	95,99	100,00	5,80	0,01	0,00	6.764,38
United Arab Emirates	98,63	100,00	5,08	0,14	0,23	7.215,22

Tabela 3.13: Agrupamento *k*-means - Dados para o Grupo 5

	ACLTC	AE	NIEEP	CER	PER	UE
Argentina	98,29	99,82	4,34	10,04	28,14	1.928,65
Aruba	86,35	100,00	3,32	6,73	14,86	1.286,32
Bolivia	62,55	91,52	4,95	17,54	31,40	636,14
Brit. Virg. Isl.	86,35	100,00	10,40	1,23	1,29	1.286,32
Cuba	79,04	99,27	2,11	19,28	3,95	1.099,50
Dominica	90,06	99,91	3,61	7,83	16,18	1.286,32
Dominican Republic	89,93	98,56	2,45	16,48	11,63	776,69
Jamaica	89,89	97,46	5,21	16,77	10,26	951,41
Mexico	85,15	99,00	3,74	9,22	15,39	1.531,76
Puerto Rico	91,00	100,00	0,41	1,84	1,84	1.286,32
St. Maarten	91,00	100,00	10,40	0,05	8,91	1.286,32
St, Kitts and Nevis	100,00	100,00	2,56	1,64	4,57	1.286,32
St, Lucia	96,97	97,36	3,19	2,13	8,91	1.286,32
St. Vinc. and Grena.	95,90	99,45	2,93	5,81	15,66	1.286,32
Turks and C, Isl.	91,00	100,00	10,40	0,57	8,91	1.286,32
Virg. Isl. (U.S.)	91,00	100,00	10,40	3,88	3,88	1.286,32
Belgium	100,00	100,00	4,74	9,20	20,80	5.539,47
Bulgaria	88,42	100,00	6,38	17,65	17,99	2.416,70
Cyprus	100,00	100,00	3,27	9,94	8,78	2.195,34
Czech Republic	97,17	100,00	5,51	14,83	11,40	4.237,90
Estonia	92,61	100,00	6,32	27,48	14,42	4.222,74
France	100,00	100,00	4,10	13,50	15,86	4.016,85
Germany	100,00	100,00	3,60	14,21	29,23	3.997,08
Greece	94,31	100,00	3,72	17,17	28,66	2.482,11
Hungary	100,00	100,00	4,32	15,56	10,58	2.568,95
Ireland	100,00	100,00	1,95	9,08	27,97	3.152,00
Italy	100,00	100,00	3,07	16,52	38,68	2.930,59
Malta	100,00	100,00	1,81	5,36	7,67	2.012,32
Netherlands	100,00	100,00	3,94	5,89	12,44	5.025,32
Poland	100,00	100,00	4,14	11,91	13,80	2.640,24
Slovak Republic	96,72	100,00	4,48	13,41	22,68	3.306,81
Slovenia	95,99	100,00	4,58	20,88	29,39	3.579,07
Spain	100,00	100,00	3,33	16,25	34,95	2.742,88
United Kingdom	100,00	100,00	3,02	8,71	24,84	3.230,62
Armenia	96,62	100,00	5,38	15,79	28,34	863,03
Azerbaijan	95,10	100,00	3,73	2,31	7,04	1.279,55
Belarus	98,01	100,00	6,47	6,77	0,82	2.900,04
Isle of Man	99,68	100,00	6,79	4,21	0,83	3.007,45

Kazakhstan	95,07	100,00	7,92	1,56	8,87	4.234,85
Kosovo	93,48	100,00	6,79	20,45	2,29	1.405,04
Moldova	91,34	100,00	8,39	14,27	5,37	984,63
Monaco	100,00	100,00	3,67	16,08	28,72	3.007,45
Russian Federation	98,15	100,00	8,41	3,30	15,86	4.819,04
San Marino	100,00	100,00	6,79	16,08	28,72	3.007,45
Serbia	75,02	100,00	6,56	21,17	26,91	2.141,06
Turkey	93,48	100,00	2,95	13,37	31,96	1.474,67
Ukraine	95,49	100,00	11,79	4,14	4,38	2.886,99
Cabo Verde	70,63	88,38	2,77	26,58	20,21	685,18
Mauritius	93,11	97,91	2,55	11,54	22,72	1.053,45
Seychelles	90,23	100,00	2,64	1,35	2,38	685,18
South Africa	83,64	85,50	8,70	17,15	2,26	2.768,09
Algeria	92,70	99,94	4,13	0,06	0,32	1.114,22
Egypt, Arab Rep,	97,31	100,00	3,51	5,71	8,26	876,54
Libya	94,41	73,24	4,21	1,97	5,09	3.353,53
Morocco	96,60	99,68	3,15	11,32	14,31	528,10
Tunisia	99,03	100,00	3,78	12,56	2,84	966,76
China	58,54	100,00	6,69	12,41	23,93	1.954,72
Hong Kong	60,44	100,00	1,49	0,85	0,28	1.946,65
Iran,	98,42	99,98	7,79	0,91	5,10	2.769,46
Iraq	97,39	99,91	3,72	0,80	3,73	1.261,22
Israel	100,00	100,00	3,59	3,71	1,89	3.042,43
Japan	100,00	100,00	3,74	6,30	15,98	3.893,27
Jordan	99,04	100,00	4,64	3,23	0,97	978,20
Korea, Rep,	96,68	100,00	6,55	2,71	1,89	5.045,49
Lebanon	94,41	100,00	4,18	3,65	2,60	1.287,77
Macao	94,41	100,00	0,66	7,05	20,41	515,20
Malaysia	96,30	100,00	4,68	5,19	9,96	2.601,45
Maldives	92,88	99,82	3,84	1,01	1,28	515,20
Singapore	100,00	100,00	2,39	0,71	1,82	5.006,62
Syrian Arab Rep,	99,05	89,81	4,03	0,52	2,31	1.013,92
Thailand	74,33	99,60	5,41	22,86	8,54	1.753,70
Turkmenistan	99,27	100,00	13,86	0,04	23,39	4.459,25
Uzbekistan	91,82	100,00	9,99	2,97	20,65	1.512,82
Yemen, Rep,	64,28	74,27	2,05	2,28	23,39	338,26
Australia	100,00	100,00	5,03	9,18	13,64	5.793,12
Marshall Isl.	64,63	92,67	11,35	11,16	0,23	1.234,30
Nauru	91,16	99,00	4,37	0,08	0,40	1.234,30
Palau	86,39	99,50	10,24	30,41	23,39	1.234,30
Tonga	58,85	96,21	3,03	1,88	5,91	1.234,30

Tabela 3.14: Agrupamento *k*-means - Dados para o Grupo 6

	ACLTC	AE	NIEEP	CER	PER	UE
Antigua and Barbuda	98,67	100,00	3,89	27,91	53,03	1286,32
The Bahamas	100,00	100,00	4,03	1,21	53,03	1.286,32
Barbados	99,41	100,00	3,78	2,79	53,03	1.286,32
Belize	85,18	91,80	5,11	35,02	45,24	1.286,32
Bermuda	86,35	100,00	2,03	2,36	53,03	1.286,32
Brazil	95,38	99,71	4,13	43,79	73,97	1.358,50
Chile	92,00	99,71	3,78	24,88	43,60	1.807,92
Colombia	91,29	98,19	2,26	23,56	68,24	689,98
Costa Rica	93,20	99,41	2,88	38,73	99,00	1.015,01
Ecuador	95,24	98,83	3,62	13,82	52,80	784,05
El Salvador	84,71	95,40	3,65	24,40	57,82	687,28
Grenada	96,50	92,58	2,97	10,92	53,03	1.286,32
Guatemala	45,06	90,57	4,48	63,65	60,39	696,79
Guyana	72,59	88,30	6,36	25,26	53,03	1.286,32
Panama	88,47	94,92	2,17	21,23	65,33	990,67
Paraguay	65,00	99,33	3,95	61,68	100,00	769,49
Peru	73,87	93,85	2,79	25,50	52,73	643,07
Suriname	89,26	95,11	3,39	24,91	60,05	1.355,63
Uruguay	97,91	99,71	3,09	58,02	88,56	1.216,64
Venezuela, RB	95,96	99,84	4,72	12,84	63,70	2.545,03
Austria	100,00	100,00	3,61	34,39	76,49	4.051,20
Croatia	92,28	100,00	4,05	33,13	66,83	2.185,45
Denmark	100,00	100,00	2,61	33,17	65,51	3.510,80
Latvia	95,05	100,00	3,91	38,10	50,17	2.148,65
Lithuania	100,00	100,00	3,86	28,96	39,41	2.275,76
Portugal	100,00	100,00	3,34	27,16	47,53	2.222,63
Romania	85,33	100,00	3,52	23,70	39,75	1.730,09
Albania	75,37	100,00	2,89	38,62	100,00	729,15
Andorra	100,00	100,00	6,79	19,75	86,12	3.007,45
Bosnia and Herzegovina	62,22	100,00	8,72	40,75	35,52	1.749,12
Georgia	76,36	99,99	5,78	28,66	78,04	824,54
Liechtenstein	100,00	100,00	3,67	63,13	96,86	3.007,45
Montenegro	68,57	100,00	4,45	43,00	49,65	1.898,29
North Macedonia	64,90	100,00	4,23	24,22	35,94	1.390,53
Switzerland	100,00	100,00	2,19	25,29	62,20	3.347,62
Gabon	78,13	89,93	6,54	82,01	43,74	3.129,08
Namibia	41,32	49,56	3,26	26,47	97,79	726,08
Afghanistan	30,10	71,50	2,46	18,42	86,05	1.357,74

Bhutan	51,75	95,65	10,41	86,90	99,99	1.357,74
Kyrgyz Republic	79,98	99,97	8,64	23,31	85,19	505,40
Tajikistan	78,55	99,65	5,01	44,66	98,47	289,07
Vietnam	63,81	100,00	5,94	35,00	36,73	669,70
New Zealand	100,00	100,00	5,42	30,79	80,08	4.225,09

REFERÊNCIAS BIBLIOGRÁFICAS

Bhattacharya, M., Paramati, S.R., Ozturk, I., Bhattacharya, S., 2016. The effect of renewable energy consumption on economic growth: evidence from top 38 countries. Applied Energy 162, 733 – 741.

Cybis, G.B., Valk, M., Lopes, S.R., 2018. Clustering and classification problems in genetics through U-statistics. Journal of Statistical Computation and Simulation 88, 1882–1902.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

IEA, 2017. Statistics & data. URL: https://www.iea.org/data-and-statistics?country=WORLD{&}fuel=Energysupply{&}indicator=Coalproductionbytype.

Kraft, J., Kraft, A., 1978. On the relationship between energy and GNP. The Journal of Energy and Development , 401–403.

Lusseau, D., Mancini, F., 2019. Income-based variation in sustainable development goal interaction networks. Nature Sustainability 2, 242–247.

Payne, J.E., 2010. A survey of the electricity consumption-growth literature. Applied Energy 87, 723 – 731.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www. R-project.org/.

The World Bank, 2019. World development indicators. URL: https://databank. worldbank.org/source/world-development-indicators.

AGRUPAMENTO DE MÚSICAS POR SIMILARIDADE

Alisson Silva Neimaier
Instituto de Matemática e Estatística - UFRGS

Amanda da Silva dos Santos
Instituto de Matemática e Estatística - UFRGS

Gabriel Fagundes da Silva
Instituto de Matemática e Estatística - UFRGS

Humberto de Lima
Instituto de Matemática e Estatística - UFRGS

Lucas da Rocha Schwengber†
Instituto de Matemática e Estatística - UFRGS

Nicolas Mathias Hahn
Instituto de Matemática e Estatística - UFRGS

RESUMO: O presente trabalho visa estudar e comparar métodos para o agrupamento de músicas com base em vetores de características extraídos das mesmas. O agrupamento é feito com o intuito de identificar padrões nas músicas de forma a criar grupos homogêneos tendo como referência critérios, como álbum e gênero. Para isso, utilizamos quatro técnicas de aprendizagem não-supervisionada: o *k*-means, o *k*-medoids, o agrupamento hierárquico aglomerativo e os modelos de mistura gaussiana. Os dados utilizados no trabalho provém do *FMA: A Dataset For Music Analysis*. Os resultados obtidos permitem avaliar e comparar os diferentes algoritmos utilizados quanto à sua capacidade de classificar as músicas em grupos alinhados com os critérios estabelecidos.
PALAVRAS-CHAVE: Música, Análise musical, Agrupamento, Aprendizado não-supervisionado

4.1 INTRODUÇÃO

Uma área de pesquisa atual e emergente que inclui diversas aplicações de Machine Learning é a área de *Music Information Retrieval*, ou MIR (Knees and Schedl, 2016). A grande gama de problemas que ela se propõe a resolver estão ligados com o princípio geral de extrair informações de músicas com o intuito, por exemplo, de gerenciar coleções de músicas para diversos fins como sistemas de recomendação, categorização, curadoria, etc. (Knees and Schedl, 2016; Futrelle and Downie, 2003).

Um problema recorrente ao armazenar diversas músicas, em uma lista de favoritos por exemplo, é posteriormente organizá-las quanto à sua similaridade. Um algoritmo capaz de realizar esta tarefa de maneira automática preservando a noção humana de similaridade

seria bastante útil. O objetivo deste trabalho é estudar, implementar e aplicar diferentes métodos de agrupamento ao problema de organização automática de músicas usando a ideia de agrupamento por similaridade. Além da implementação de um algoritmo que realize esta tarefa, também temos o objetivo de comparar diferentes métodos de agrupamento, seguindo uma formulação matemática especificamente desenvolvida para o problema.

Embora um dos desenvolvimentos centrais do MIR seja a classificação automática de gêneros musicais, utilizar classificação para resolver o problema proposto neste trabalho não é uma alternativa viável. Em primeiro lugar, isto nos restringe a apenas uma noção de similaridade entre músicas: a de gênero. Dependendo do contexto isto pode nos levar a agrupamento triviais, por exemplo se todas as músicas que queremos organizar são do mesmo gênero mas são de artistas diferentes. Um segundo problema com esta abordagem é a necessidade de se ter um número massivo de *rótulos* à disposição e, consequentemente, observações para cobrir todos os diferentes gêneros que podem ser encontrados em aplicações. O fato de os algoritmos de agrupamento não necessitarem de rótulos externos pré-definidos, os torna especialmente úteis neste contexto em que não é viável passar todas as informações "a priori" do que o algoritmo pode vir a encontrar.

4.2 METODOLOGIA

4.2.1 Formulação Matemática

Matematicamente, o problema que estamos interessados em resolver pode ser descrito da seguinte forma. Dadas N músicas, gostaríamos de agrupá-las em *k* grupos distintos. Se a cada música correspondermos um vetor de *p* características extraídas do seu respectivo arquivo de áudio, temos essencialmente um problema de agrupamento. O objetivo é agrupar as *N* músicas observadas com suas características, denotadas por $\boldsymbol{m}_i = (x_{i1}, \cdots, x_{ip}), i = 1, \cdots, \mathrm{N}$, em k grupos disjuntos de forma a satisfazer um determinado critério que capture o objetivo desta aplicação: cada grupo deverá conter músicas o mais similares possível segundo alguma noção de similaridade que seja útil para seres humanos. Para avaliar quão perto estamos do objetivo, é necessário especificar um critério de avaliação dos agrupamentos. Por simplicidade, utilizaremos a notação M_i para indicar a música i enquanto arquivo de áudio/a entidade que ouvimos, e $\boldsymbol{m}_i$ para indicar o vetor de características extraído desta música.

4.2.2 Critério de avaliação

No âmbito geral, uma maneira de construir um critério é a seguinte: supomos que para cada par de músicas (M_i, M_j) existe uma distância subjetiva d_{ij} que é percebida por um indivíduo ao ouví-las e compará-las. Se $M_i = M_j$, então $d_{ij} = 0$. Fora isso, esta distância não é algo a que temos acesso (nem algo que está muito bem definido), mas é possível intuir como

ela seria ordenada e que tipos de agrupamentos fazem algum sentido para a percepção auditiva humana. Por exemplo, se selecionarmos uma música dos Rolling Stones M_1 = *Jumpin' Jack Flash*, uma música dos Beatles M_2 =*And Your Bird Can Sing* e uma fuga de Bach M_3 = *Fuga em Sol menor (BWV 578)*, é razoável supor que $d_{13} > d_{12}$ e $d_{23} > d_{12}$ de forma que o agrupamento mais razoável em dois grupos seria $\{\{M_1, M_2\}, \{M_3\}\}$, pois desta forma as músicas mais "parecidas" estão no mesmo grupo e separadas da mais "distante".

Para fins práticos, escolhemos uma forma de aproximar esta distância e torná-la bem definida, determinando uma distância entre músicas que capture pelo menos parte de nossa intuição. Sendo assim, uma escolha natural é definir uma distância que determina se duas músicas possuem pelo menos um *metadado* categórico em comum. Por exemplo, se são do mesmo artista, do mesmo álbum, do mesmo gênero, etc. Fixado um atributo $A(M_i)$ definimos esta distância como:

$$d_{ij} = \begin{cases} 0, & \text{se } A(M_i) = A(M_j); \\ 1, & \text{se } A(M_i) \neq A(M_j). \end{cases}$$

Retomando o exemplo citado das 3 músicas, se tomarmos M_1 = *Jumpin' Jack Flash*, M_2 = *And You Bird Can Sing* e M_3 = *Fuga em Sol menor (BWV 578)*, e $A(M_i)$ como sendo o gênero da música i, temos que $A(M_1) = A(M_2)$ = *Rock* e $A(M_3)$ = *Música Clássica* e portanto $d_{12} = 0$ e $d_{13} = d_{23} = 1$ que dá origem à ordem que intuímos.

4.2.3 A métrica de desempenho

Com uma medida de distância em mãos, podemos definir uma métrica, no sentido de uma estatística de desempenho do agrupamento, que capture a ideia de quão bom é um agrupamento. Optamos por usar uma métrica especialmente feita com a aplicação em mente. Ela é construída da seguinte maneira: suponha que um usuário execute um algoritmo de agrupamento hipotético em uma pasta de músicas as quais só é possível diferenciar ouvindo. O algoritmo devolverá um agrupamento $C = \{C_1, \cdots, C_k\}$ (onde os C_i's podem ser interpretados como as pastas). Para verificar se o algoritmo funciona, o usúario escolhe uma das pastas (de maneira uniforme já que o usúario não possui nenhuma informação a priori sobre o tamanho das pastas). Dentro desta pasta, o usuário então seleciona aleatoriamente duas músicas (sem reposição) e avalia mentalmente a "distância" subjetiva entre elas.

Podemos entender o resultado deste experimento como uma variável aleatória D que representa um custo (se a pasta contém apenas uma música, convencionamos que $D = 0$). Ou seja, se D é grande (comparado com uma expectativa de distância média entre as músicas) o usuário tenderá a crer que o algoritmo não funciona adequadamente, pois o algoritmo agrupou músicas "distantes" em uma mesma pasta. Nesse contexto, dentre diferentes métodos de agrupamento, escolhemos o que minimiza $\mathbb{E}(D)$ que, portanto,

funcionará como uma função perda. Na prática a variável aleatória D assume os valores da distância que escolhemos para "aproximar" a distância subjetiva avaliada pelo usúario.

Note que, com a distância indicadora, $\mathbb{E}(D)$ nada mais é do que a probabilidade de, ao abrir uma pasta aleatória e selecionar duas músicas, o evento $[\mathrm{A}(M_i) \neq \mathrm{A}(M_j)]$ ocorrer. Dessa forma, quanto mais variados os atributos das músicas de um grupo, maior será esta probabilidade e, consequentemente, a função perda, o que se reflete em um desempenho ruim.

Dada uma música M_i, denotaremos por G_i o gênero desta música. Para as análises realizadas na próxima seção, utilizaremos a distância indicadora do gênero, que é 1 se as duas músicas são de gêneros distintos e 0 se forem do mesmo gênero. Neste caso, supondo que temos L gêneros distintos e K grupos, a função perda possui a seguinte forma:

$$
\begin{aligned}
\mathbb{E}(D) = \mathbb{E}(\mathbb{E}(D|F_k)) = \mathbb{E}\left(\mathbb{P}\left(G_i \neq G_j|F_k\right)\right) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{P}\left(G_i \neq G_j|F_k\right) \\
= \frac{1}{K}\sum_{k=1}^{K} \frac{\#\text{Pares de Músicas de Gêneros Distintos em } F_k}{\#\text{Pares de Músicas em } F_k} \\
= \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} \frac{C_{ik}C_{jk}}{\binom{C_k}{2}}, \qquad (4.1)
\end{aligned}
$$

onde F_k denota o grupo escolhido aleatoriamente pelo usúario, C_{ik} é o número de músicas do gênero i no grupo k e k e $C_k = \sum_{i=1}^{L} C_{ik}$ é o número total de músicas no grupo k.

Na Figura 4.1 apresentamos um histograma dos valores da função perda assim definida após 10.000 agrupamentos aleatórios em dois conjunto de dados contendo respectivamente n = 5336 e n = 2729 músicas de k = 3 e k = 5 gêneros. O número de grupos escolhido foi igual ao de gêneros do respectivo conjunto de dados. Cada agrupamento aleatório consiste em uma alocação das músicas nos grupos através de um sorteio uniforme de um grupo para cada música. Analisando a Figura 4.1 é possível ter uma ideia da distribuição de valores da função perda (4.1) para um agrupamento aleatório. Esta distribuição empírica foi usada como base para determinar quao significativo é um agrupamento realizado por algum método comparado com um realizado de maneira aleatória. É interessante notar o aumento na média da função perda conforme aumenta-se a variedade de gêneros musicais considerados, indicando um problema mais desafiador.

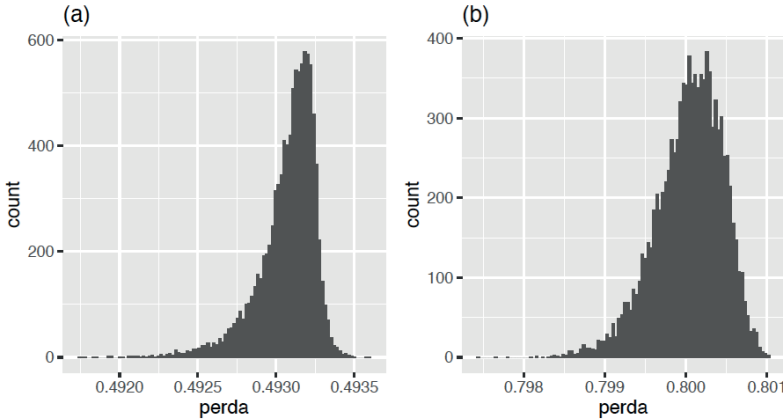



Figura 4.1: (a) Histograma de valores da função perda para 10.000 agrupamentos aleatórios em k = 3 grupos (sorteio uniforme de qual grupo cada música pertence) num conjunto de dados de n = 5336 músicas de 3 gêneros. (b) Histograma de valores da função perda para 10.000 agrupamentos aleatórios em k = 5 grupos (sorteio uniforme de qual grupo cada música pertence) num conjunto de dados de n = 2729 músicas de 3 gêneros.

4.2.4 Metódos clássicos de Redução de dimensionalidade e pré-processamento

Utilizaremos dois métodos clássicos de redução de dimensionalidade e pré-processamento dos dados. Por sua relevância, ambos serão introduzidos agora.

PCA

A análise de componentes principais, também conhecida como PCA (do inglês Principal Component Analysis) é uma abordagem não supervisionada que envolve n observações de um conjunto de variáveis $X_1, \cdots, X_p$ sem uma variável resposta associada. O objetivo do PCA é obter as melhores aproximações lineares (através de hiperplanos afins em dimensões $1, \cdots, p$) para os dados observados capturando a maior variação possível (veja a seção 14.5 de Hastie et al., 2009).

Cada componente principal é dada por uma combinação linear das variáveis $X_1, \cdots, X_p$ da seguinte forma:

$$\begin{aligned} Z_1 &= \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p \\ Z_w &= \phi_{12}X_1 + \phi_{22}X_2 + \cdots + \phi_{p2}X_p \\ &\vdots \\ Z_q &= \phi_{1q}X_1 + \phi_{2q}X_2 + \cdots + \phi_{pq}X_p \end{aligned} \quad (4.2)$$

com as seguintes restrições:

1. $Z_1, \cdots, Z_q$ são não correlacionados, $q \leq p$;

2. $\sum_{j=1}^{p} \phi_{ji}^2 = 1, \ i \in \{1, \cdots, q\}$.

Fazendo a decomposição em valores singulares da matriz de design associada ao problema (com média das colunas igual a zero), $X = UDV'$, é possível mostrar que os coeficientes ϕ_{ij} que resolvem (4.2) com as restrições impostas são dados pelas entradas da matriz U até a coluna q (Hastie et al., 2009). Neste trabalhos foi utilizada uma implementação em Python, através do pacote sklearn (Pedregosa et al., 2011).

ICA

Uma motivação clássica para o ICA é o famoso *cocktail party problem* que consiste em resgatar as fontes de som (conversas independentes) a partir dos registros sonoros das conversas misturadas captados por microfones dispostos na sala em diferentes locais (Hyvärinen and Oja, 2000). Diferentemente do PCA, o ICA busca uma representação das variáveis do banco através de uma combinação linear de variáveis latentes independentes (não apenas não-correlacionadas) (Hastie et al., 2009; Hyvärinen and Oja, 2000). Na prática, podemos dizer que o PCA ajuda a reduzir a dimensão dos dados enquanto o ICA ajuda a separá-los em suas componentes constituintes. Um vetor aleatório $\boldsymbol{X} = (X_1, \cdots, X_p)$ é escrito no ICA como uma combinação linear de variáveis aleatórias $S_1, S_2, \cdots, S_p$:

$$\begin{aligned} X_1 &= \mathbf{a}_{11}S_1 + \cdots + \mathbf{a}_{1p}S_p \\ X_2 &= \mathbf{a}_{21}S_1 + \cdots + \mathbf{a}_{2p}S_p \\ &\vdots \\ X_p &= \mathbf{a}_{p1}S_1 + \cdots + \mathbf{a}_{pp}S_p, \end{aligned} \tag{4.3}$$

Denotando $\boldsymbol{S} = (S_1, \cdots, S_p)$ isto pode ser reescrito matricialmente como $\boldsymbol{X} = \boldsymbol{AS}$, onde $\boldsymbol{A}$ é uma matriz quadrada que gostaríamos de estimar de forma que $\boldsymbol{S} = \mathbf{A}^{-1}X$ satisfaça:

1. $S_1, \cdots, S_p$ são independentes;
2. $S_1, \cdots, S_p$ têm distribuição não-Gaussiana;
3. $\text{Var}(S_i) = 1, i \in \{1, \cdots, p\}$.

A menos do sinal de S_i, a solução para este problema está bem definida, embora os métodos para estimá-la são bem mais complexos do que o caso do PCA. Além disso, note que para fazer redução de dimensionalidade é necessário flexibilizar esta hipótese de que $\mathbf{A}$ é uma matriz quadrada, pois deseja-se usar menos componentes independentes S_i do que características X_i. Neste caso, é necessário considerar uma versão levemente modificada do problema. Mais detalhes de obtenção computacional e da formulação podem

ser encontrados em Hyvärinen and Oja (2000). O algoritmo clássico para estimar a matriz **A** é chamada de *FastICA*. Neste trabalhos foi utilizada uma implementação deste tanto em R, através do pacote fastICA, quanto em Python, através do pacote sklearn (Pedregosa et al., 2011).

4.2.5 Extração de Características

A outra abordagem utilizada para o pré-processamento dos dados é baseada nas representações utilizadas por certos algoritmos de classificação. Em linhas gerais, treina-se um algoritmo de classificação para classificar, por exemplo, o gênero das músicas, e olha-se quais variáveis, ou combinações delas, são as mais relevantes para o algoritmo. A ideia é que isto dá uma indicação de quais características dos dados capturam melhor a noção de gênero, por exemplo.

Redes Neurais

Dentre os métodos de classificação que envolvem algum tipo de aprendizado de representação dos dados, as *redes neurais* é possivelmente um dos mais famosos. Uma motivação clássica para a arquitetura em camadas de uma rede neural "feedforward" é a noção de cada camada "aprende" uma representação dos dados em um nível de abstração cada vez maior. Portanto, um método "natural" de representação dos dados induzido por essa arquitetura é a saída da penúltima camada da rede. A ideia é que uma rede neural que consegue classificar bem os gêneros deve, em teoria, conter uma representação dos dados em sua última camada capaz de discernir bem os mesmos, colocando músicas de gêneros parecidos próximas, e músicas de gêneros diferentes distantes. A implementação utilizada para redes neurais foi a do pacote sklearn (Pedregosa et al., 2011) do Python.

Árvores de Decisão

Para utilizar as Árvores de decisão como um método de seleção de variáveis, treinamos um classificador por métodos de Monte Carlo. Foram feitas 1.000 replicações de amostras de tamanho 1.000 do banco de dados e em cada uma das replicações tomamos quais variáveis foram consideradas importantes. Aquelas consideradas importantes em mais de 60% das replicações foram selecionadas.

4.2.6 Métodos de Agrupamento

k-means

O algoritmo *k*-means é uma técnica de agrupamento caracterizada por não definir previamente os padrões que serão gerados, e pela dependência de uma entidade externa

que informe qual a quantidade *k* de grupos a ser formada e, em seguida, utilizar uma técnica de realocação iterativa baseada em similaridade. A medida de similaridade aplicada aos objetos é expressa como uma função, ou métrica, que mede a distância do objeto aos centróides dos *k* grupos gerados (Hastie et al., 2009).

Ao longo da análise, foi possível verificar algumas desvantagens deste método. A mais notória delas é a sensibilidade à partição inicial, gerada pela escolha aleatória dos centróides. Por conta disso, foram usadas variações da função padrão do R com o intuito de tornar o método mais estável. Para análise do *k*-means foram usadas algumas funções de pacotes do R e outras desenvolvidas pelos autores e disponíveis em https://github.com/Lucas-Schwengber/Music-Clust-Dados-Testes.git. Estas são descritas abaixo:

- perda_gen: define a métrica da perda que será utilizada (função própria).
- matrizdeconfusao: encontra a matriz de confusao, organizando as categorias corretamente (função própria).
- SeedTestAccuracy_kmeans: executa o *k*-means 100 vezes para testar o quanto o resultado depende da semente inicial (função própria).
- kmeans: aplica o método de agrupamento *k*-means (função base do R)
- KMeans_rcpp: aplica o método de agrupamento *k*-means com uma semente fixa (do pacote ClusterR - veja Mouselimis, 2021)

k-medoides

O *k*-medoides é um método de agrupamento para particionar um conjunto de dados em *k* grupos. Embora sendo relacionado ao *k*-means, o *k*-medoides não calcula o seu centro de acordo com a média de todos os pontos do grupo. Assim, cada grupo é representado pelo ponto no qual a diferença média entre ele e todos os outros membros do grupo é mínima. Esse ponto é denominado medoide e ele necessariamente pertence ao grupo que representa. Geralmente, é um método mais estável que o *k*-means, pois minimiza uma soma de dessemelhanças gerais em pares, em vez de uma soma de distâncias euclidianas quadradas, tornando o método menos sensível à presença de outliers.

Para a aplicação do método, foi usado o algoritmo PAM (*Partitioning Around Medoids*), que baseia-se na busca de *k* objetos representativos no conjunto de dados, e no cálculo da função objetivo com o intuito de encontrar os medoides que minimizam a soma das diferenças entre as observações e o objeto representativo mais próximo.

Agrupamento Hierárquico Aglomerativo

Para encontrar grupos similares dentro de um banco de dados com *n* observações de $X_1, \cdots, X_p$, o método de agrupamento hierárquico aglomerativo segue os seguintes passos:

1. Considera-se cada uma das *n* observações como um grupo unitário;

2. Os dois grupos mais similares são aglomerados, formando um único grupo;

3. Calcula-se a similaridade deste novo grupo com os outros;

4. Repete-se os passos 2 e 3 até que todas as observações formem um grupo de tamanho *n*.

Dentre as vantagens de utilizarmos o agrupamento hierárquico destacamos que o método não exige que o número de grupos finais seja definido de antemão e também a sua fácil representação gráfica através de dendogramas. A similaridade entre grupos pode ser definida de diversas formas. Dentre as opçoes oferecidas pelo algoritmo utilizado, a que mostrou melhores resultados foi a “ward.D2”, que aplica o método de Ward conforme Ward (1963) (veja também Murtagh and Legendre, 2011).

Método de Misturas Gaussianas

O modelo de mistura gaussiana é um caso particular de uma classe de modelos denominada *modelos de mistura* (*mixture models*, Murphy, 2012, Capítulo 11). Neste tipo de modelo, pensamos na distribuição do vetor de características$= (X_1, \cdots, X_p)$ como sendo gerado em duas etapas:

1. Sorteamos um valor z_k para uma variável aleatória latente discreta *Z* que assume valores $k \in \{1, \cdots, K\}$ com probabilidade $\pi_k = \mathbb{P}(Z = k)$.

2. Sorteamos um valor para $\boldsymbol{X}$ de acordo com uma distribuição $p(\boldsymbol{x}|Z = k) = p_k(\boldsymbol{x})$

Supondo que o modelo possui um conjunto de parâmetros *θ*, o procedimento acima dá origem à seguinte forma da distribuição de $\boldsymbol{X}$ (seja ela discreta ou contínua):

$$f(\boldsymbol{x}|\theta) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{x}|\theta_k)$$

No caso particular do modelo de mistura gaussiana supomos, que $\boldsymbol{X}|Z = k$ tem distribuição normal *p*-variada com média $\mu k \in \mathbb{R}^p$, matriz de variância-covariância Σ_k e densidade $\phi_p(\cdot\, ; \mu k, \Sigma_k)$, de forma que a densidade de $\boldsymbol{X}$ dado θ assume a forma:

$$f(\boldsymbol{x}|\theta) = \sum_{k=1}^{K} \pi_k \phi_p(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k) \qquad (4.4)$$

Note que uma maneira natural de pensar na variável Z é como uma indicadora do grupo a partir do qual $\boldsymbol{X}$ foi gerado. Por exemplo, na aplicação em questao podemos pensar na distribuição condicional de $\boldsymbol{X}|Z = k$ como a distribuição que o vetor de características $\boldsymbol{X}$ segue dado que ele foi extraído de uma música do gênero *k*. Desta maneira, assumindo que os dados de fato seguem o modelo (4.4), podemos utilizar a regra de Bayes para obter a probabilidade de o gênero ser *k* dada uma observação $\boldsymbol{x}$,

$$p(Z=k|\boldsymbol{x},\theta)=\frac{f(\boldsymbol{x}|Z=k,\theta)p(Z=k|\theta)}{\sum_{j=1}^{K} f(\boldsymbol{x}|Z=j,\theta)p(Z=j)}.$$

Com isto, podemos obter um agrupamento das observações associando a cada observação $\boldsymbol{x}_i$ o grupo Z_i^* que maximiza $p(Z=k|\boldsymbol{x},\theta)$, isto é,

$$Z_i^* = \text{argmax}_k\{\log(f(\boldsymbol{x}|Z=k,\theta)) + \log(p(Z=k|\theta))\} \quad (4.5)$$

A utilização de (4.5) requer ajustar o modelo (4.4) utilizando os dados observados $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$. Fixado *K*, (4.4) pode ser estimada utilizando o algoritmo EM (veja a seção 11.4.2 de Murphy, 2012). Este método está implementado no pacote sklearn (Pedregosa et al., 2011) do Python.

4.2.7 Algumas hipóteses

A primeira etapa do trabalho consistirá na comparação do desempenho de diferentes métodos de agrupamento para dados de músicas das quais conhecemos vários *metadados*, cenário no qual podemos comparar o desempenho dos algoritmos usando a função perda. Posteriormente, os algoritmos com melhor desempenho na primeira etapa foram utilizados em cenários que visam simular aplicações reais. Para que estas duas partes estejam interligadas, são necessárias as seguintes suposições:

- Os testes no banco de dados da Seção 4.4 trazem informações a respeito de quais algoritmos apresentariam melhor desempenho nos cenários em que não é possível calcular o valor da função perda.
- A escolha do número de grupos não é relevante para avaliação do algoritmo pois cabe ao usuário. Isto é importante pois a métrica proposta não penaliza a quantidade de grupos.
- As características escolhidas para representar as músicas contém informação suficiente para capturar a noção de distância escolhida. Ou seja, um algoritmo perfeito conseguiria, com base somente nas características, resgatar a noção de proximidade subjetiva entre músicas que estamos buscando (a menos das limitações subjetivas presentes na definição desta distância). Esta hipótese é crucial para que o agrupamento seja possível pois, na prática, as únicas informações disponíveis são as características.

4.3 BANCO DE DADOS

O banco utilizado para extrair os dados para os testes de desempenho dos métodos foi o FMA: A Dataset For Music Analysis (Defferrard et al., 2017, https://github.com/mdeff/ fma). O banco completo contém 106.574 músicas sob a *Creative Commons-Licence* divididas em 161 gêneros, cada música podendo ter mais de um gênero. Usou-se porém uma versão resumida dos dados que contém diversos *metadados* das músicas como, artista, ano,

gênero(s), e contendo mais de 500 características numéricas que descrevem cada música. Dentre os conjuntos de 500 características, as escolhidas para análise foram um conjunto de estatísticas dos Mel Frequency Cepstral Coefficients (MFCC) (Logan, 2000; Rabiner et al., 1993). Para evitar ambiguidades, foram considerados apenas músicas que possuíam apenas um gênero.

MFCC

Os Mel Frequency Cepstral Coefficients (MFCC) são coeficientes que extraem características de sinais áudios de maneira a tentar simular a percepção auditiva humana. Tradicionalmente são usados em tarefas de reconhecimento de voz (Rabiner et al., 1993; Logan, 2000), mas podem funcionar também para dados musicais (Logan, 2000). Um tutorial acessível mas detalhado de sua implementação pode ser encontrado no site da Practical Cryptography (https://tinyurl.com/442fzsut). As analises apresentadas neste trabalho foram feitas em python utilizando a biblioteca librosa (McFee et al., 2019).

Os MFCC propriamente ditos consistem em 20 coeficientes (em algumas aplicações usa-se mais, outras menos) extraídos de cada *frame* (pequenos trechos encadeados da série temporal) do áudio, geralmente a uma taxa de 22050 kHz. Dada esta alta taxa de amostragem, é fácil ver que a dimensão dos dados ficaria muito grande se usássemos todos os 20 coeficientes de cada um dos frames. Por essa razão, foram utilizadas apenas algumas estatísticas que resumem a distribuição dos MFCCs ao longo dos frames, com informações sobre o mínimo, máximo, média, mediana, desvio padrão, curtose e assimetria de cada um dos 20 coeficientes (7 estatísticas da distribuição de 20 coeficientes, totalizando 140 características).

4.4 COMPARAÇÃO DOS MÉTODOS DE AGRUPAMENTO

4.4.1 Conjunto de dados utilizado

Para comparar a eficácia dos possíveis métodos de agrupamento considerados na Seção 4.2.6 para solucionar o problema proposto, foi escolhido um subconjunto do banco de dados apresentado na Seção 4.3 contendo três gêneros: *Hip-Hop*, *Música Clássica* e *Old-Time/Historic* (músicas populares do início do século 20). Com isto foi obtido um banco com 5.336 observações das 140 características sendo 3.552 delas de *Hip-Hop*, 1.230 de *Música Clássica* e 554 de *Old-Time/Historic*. Os dados podem ser encontrados no repositório: https://github.com/Lucas-Schwengber/Music-Clust-Dados-Testes.git. Nas Figuras 4.2 e 4.4.1 apresentamos uma visualização dos dados em termos dos três primeiros componentes independentes e os três primeiros componentes principais, respectivamente. Denominaremos este subconjunto dos dados como *Banco Hip-Cla-Old*.

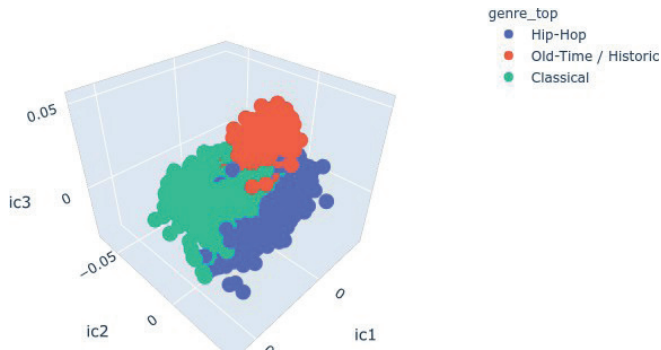



Figura 4.2: Visualização o subconjunto dos dados usando 3 componentes independentes.

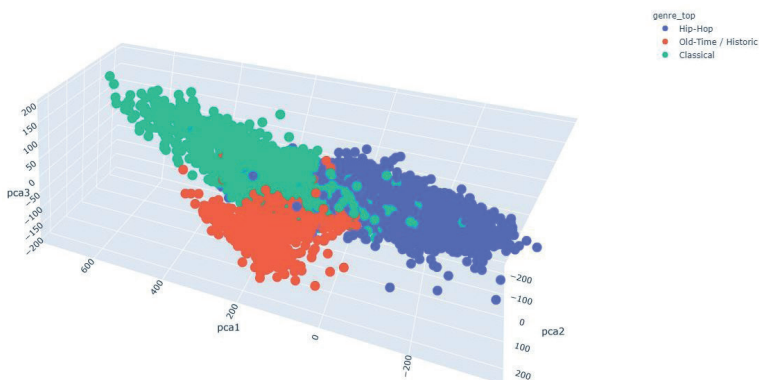



Figura 4.3: Visualização do subconjunto dos dados usando as 3 primeiras componentes principais.

4.4.2 Os Cenários Testados

Foram testados os 4 métodos de agrupamento apresentados na Seção 4.2.6 com e sem pré-processamento via PCA e ICA nos dados do *Banco Hip-Cla-Old*. Os métodos foram aplicados por amostragem. Amostramos 1.000 músicas do *Banco Hip-Cla-Old* 1.000 vezes, aplicamos em cada amostra os métodos de agrupamento com e sem pré-processamento via PCA e ICA e para cada agrupamento obtido, calculamos a função perda usando a métrica indicadora de gênero. Os resultados apresentados resumem a distribuição dos valores da função perda observados nas 1.000 reamostragens. Além disso, para cada tipo de agrupamento foram utilizados $k = 3, 5$ e 7 grupos.

Para os agrupamentos *k*-means, *k*-medoides e hierárquico foram testados 6 tipos de vetores de entrada, com um tipo a mais no hierárquico, a saber:

- As 140 características brutas do MFCC (coluna Dados);
- PCA com 3 componentes das 140 características brutas do MFCC (coluna PCA 3V);

- PCA com o menor número de componentes que capture 60% da variância das 140 características brutas do MFCC (coluna PCA 60%);
- PCA com o menor número de componentes que capture 70% da variância das 140 características brutas do MFCC (coluna PCA 70%);
- PCA com o menor número de componentes que capture 80% da variância das 140 características brutas do MFCC (coluna PCA 80%);
- PCA com o menor número de componentes que capture 90% da variância das 140 características brutas do MFCC (coluna PCA 90%).
- As características dos dados brutos do MFCC normalizadas que foram consideradas importantes numa aplicação do algoritmo de árvores do *rpart* (ver 4.2.5 (coluna Trees).

Os resultados são apresentados nas Tabelas 4.1 até 4.3.

Tabela 4.1: Estatísticas descritivas dos valores observados da função perda utilizando o método *k*-means.

			$k = 3$			
	Dados	PCA 3V	PCA 60%	PCA 70%	PCA 80%	PCA 90%
Média	0,249	0,273	0,249	0,246	0,246	0,250
Mediana	0,261	0,269	0,261	0,260	0,260	0,263
D.P.	0,068	0,035	0,065	0,069	0,069	0,069
Min	0,062	0,138	0,074	0,059	0,062	0,060
Max	0,543	0,399	0,525	0,385	0,420	0,535
			$k = 5$			
Média	0,135	0,185	0,134	0,128	0,132	0,133
Mediana	0,122	0,185	0,124	0,117	0,120	0,120
D.P	0,045	0,041	0,039	0,038	0,041	0,044
Min	0,060	0,095	0,064	0,058	0,063	0,055
Max	0,411	0,313	0,346	0,297	0,356	0,357
			$k = 7$			
Média	0,124	0,158	0,130	0,124	0,124	0,125
Mediana	0,118	0,153	0,126	0,121	0,119	0,118
D.P.	0,036	0,035	0,031	0,030	0,034	0,034
Min	0,059	0,081	0,069	0,061	0,054	0,053
Max	0,311	0,287	0,251	0,238	0,326	0,253

Tabela 4.2: Estatísticas descritivas dos valores observados da função perda utilizando o método das *k*-Medoides.

			k = 3			
	Dados	PCA 3V	PCA 60%	PCA 70%	PCA 80%	PCA 90%
Média	0,262	0,246	0,250	0,263	0,268	0,268
Mediana	0,267	0,246	0,258	0,262	0,269	0,270
D.P.	0,055	0,018	0,048	0,038	0,041	0,047
Min	0,082	0,164	0,073	0,083	0,054	0,077
Max	0,415	0,327	0,397	0,395	0,413	0,414
			k = 5			
Média	0,153	0,173	0,136	0,134	0,142	0,148
Mediana	0,144	0,174	0,133	0,131	0,137	0,141
D.P	0,042	0,029	0,028	0,029	0,034	0,039
Min	0,069	0,095	0,065	0,068	0,067	0,065
Max	0,327	0,255	0,244	0,243	0,284	0,327
			k = 7			
Média	0,138	0,138	0,124	0,128	0,130	0,135
Mediana	0,138	0,137	0,124	0,128	0,128	0,135
D.P.	0,026	0,025	0,022	0,024	0,024	0,026
Min	0,071	0,075	0,058	0,059	0,062	0,066
Max	0,258	0,232	0,189	0,206	0,228	0,263

Tabela 4.3: Estatísticas descritivas dos valores observados da função perda utilizando o método Hierárquico.

				k = 3			
	Dados	PCA 3V	PCA 60%	PCA 70%	PCA 80%	PCA 90%	Trees
Média	0,148	0,247	0,170	0,156	0,152	0,149	0,105
Mediana	0,124	0,239	0,146	0,130	0,128	0,125	0,101
D.P.	0,070	0,054	0,074	0,071	0,071	0,072	0,034
Min	0,045	0,089	0,056	0,051	0,042	0,044	0,033
Max	0,424	0,411	0,405	0,388	0,373	0,389	0,300
				k = 5			
Média	0,126	0,179	0,139	0,130	0,128	0,126	0,106
Mediana	0,121	0,177	0,134	0,125	0,124	0,122	0,101
D.P.	0,036	0,040	0,035	0,034	0,035	0,036	0,035
Min	0,042	0,085	0,042	0,052	0,042	0,044	0,029
Max	0,302	0,327	0,280	0,245	0,276	0,273	0,250
				k = 7			
Média	0,128	0,169	0,139	0,133	0,131	0,129	0,106
Mediana	0,125	0,167	0,138	0,131	0,127	0,125	0,103
D.P.	0,034	0,034	0,031	0,031	0,033	0,033	0,031
Min	0,042	0,091	0,040	0,060	0,054	0,049	0,032
Max	0,249	0,312	0,250	0,251	0,251	0,248	0,236

Para o agrupamento por misturas gaussianas foram testados 5 tipos de vetores de entrada:

- 140 características brutas do MFCC (coluna Dados);
- ICA com 3 componentes das 140 características brutas do MFCC (coluna ICA 3V);
- PCA com 3 componentes das 140 características brutas do MFCC (coluna PCA 3V);
- PCA com o menor número de componentes que capture 90% da variância das 140 características brutas do MFCC (coluna PCA 90%).
- A saída da camada escondida de uma rede neural com uma única camada escondida com 370 neurônios treinada nas observações do banco para classificar gêneros musicais (coluna Neural).

Os resultados são apresentados na Tabela 4.4.

Tabela 4.4: Estatísticas descritivas dos valores observados da função perda utilizando o método das Misturas Gaussianas.

			$k = 3$		
	Dados	ICA 3V	PCA 3V	PCA 90%	Neural
Média	0,254	0,162	0,248	0,265	0,148
Mediana	0,251	0,138	0,244	0,259	0,108
D.P.	0,022	0,063	0,051	0,055	0,086
Min	0,196	0,072	0,094	0,090	0,043
Max	0,360	0,373	0,362	0,378	0,386
			$k = 5$		
Média	0,142	0,144	0,120	0,114	0,166
Mediana	0,147	0,137	0,115	0,108	0,148
D.P.	0,043	0,033	0,029	0,031	0,066
Min	0,059	0,072	0,064	0,052	0,059
Max	0,281	0,273	0,252	0,279	0,414
			$k = 7$		
Média	0,114	0,145	0,122	0,115	0,165
Mediana	0,111	0,143	0,121	0,113	0,154
D.P.	0,026	0,033	0,027	0,026	0,052
Min	0,061	0,057	0,062	0,050	0,048
Max	0,228	0,293	0,215	0,232	0,406

4.4.3 Comparação

Para k = 3, os valores da função perda ficam torno de 0,148 a 0,273 o que, embora relativamente distante de 0, se olharmos para a distribuição da função perda para o caso de 3 gêneros Figura 4.1, vemos que um agrupamento aleatório gera valores concentrados em 0,49 com mínimo superior a 0,48 para 10.000 iterações, ou seja, os algoritmos apresentam um desempenho significativamente melhor do que um agrupamento aleatório. Isto corrobora a hipótese de que as características usadas contém informação sobre a similaridade que queremos emular e os métodos de agrupamento usados são capazes de extrair parte dessa informação.

Para k = 5 os valores da média ficam entre 0,114 e 0,185. Isto representa uma queda representativa em quase todos os métodos comparado com o caso k = 3, com exceção do agrupamento hierárquico. Para k = 7 os valores ficam entre 0,114 e 0,169 que não representa uma queda tão significativa comparado com o caso k = 5.

Quanto ao melhor método, se olharmos para as médias, para k = 3 o agrupamento hierárquico tende a se sobressair. O Agrupamento por mistura gaussiana, o k-means e o k-Medoides apresentaram um desempenho similar entre si. O pré-processamento via PCA e ICA parece ter melhorado os resultados apenas para o agrupamento por mistura gaussiana. Para os demais métodos, este pré-processamento em geral piora o desempenho. O uso da saída da última camada de uma rede neural teve um desempenho melhor do que os outros métodos de pré-processamento apenas para k = 3. Já a extração via árvores tem um desempenho consistentemente melhor para todos os k's.

Embora exista uma tendência de a função perda diminuir quando se aumenta o número de grupos, note que para uma aplicação real envolvendo um banco com 1.000 músicas, as 3 opçoes de tamanho de grupos, k = 3, 5 e 7 representam um número muito pequeno de opções para se obter uma organização útil. Para obter pastas com 30 músicas em média, é necessário que ao menos k = 33, o que representa um número de grupos bem maior que os testados. Supondo que a tendência de queda da função perda a medida que k aumenta, observada nos casos considerados, se mantenha, podemos esperar que a função perda destes métodos em aplicações reais tenderá a ser ainda menor.

De maneira geral é possível concluir que o método de agrupamento mais complexo (mistura gaussiana) e o pré-processamento (ICA ou PCA) não melhoraram o desempenho da função perda de maneira representativa. Portanto é natural optar por usar os métodos mais simples, em especial o agrupamento hierárquico, e nas características brutas do *MFCC* sem realizar pré-processamento.

4.5 SIMULANDO APLICAÇÕES

Os resultados obtidos na seção anterior foram baseados em um banco com um grande número de músicas com informações *a priori* em abundância acerca das músicas, como por exemplo, o gênero. Este cenário é um pouco distante de aplicações reais. Nesta seção apresentamos resultados em dados que visam simular aplicações reais.

Para realizar estes testes foi implementada uma rotina em Python que, dada uma pasta com músicas, esta extrai as mesmas características usadas no banco de teste e roda um algoritmo de agrupamento nelas e, com base neste agrupamento, organiza as músicas em *k* pastas. Foram testados diferentes métodos de agrupamento em dois cenários que visam simular aplicações de diferentes graus de dificuldade.

4.5.1 1º Cenário - Mistura de Álbuns

No primeiro cenário, foram misturados 3 álbuns de estilos musicais distintos em uma mesma pasta. Os três álbuns escolhidos foram:

- *Revolver - The Beatles* (14 músicas);
- *"The Black Album" - Metallica* (12 músicas);
- *Norman F* Rockwell - Lana Del Rey* (14 músicas);

O objetivo foi tentar diferentes métodos de agrupamento para separar o conjunto de músicas em três pastas resgatando os três álbuns originais. Dos diferentes métodos testados, dois estão ilustrados abaixo, um dos quais foi capaz de separar perfeitamente as músicas segundo o critério "álbum".

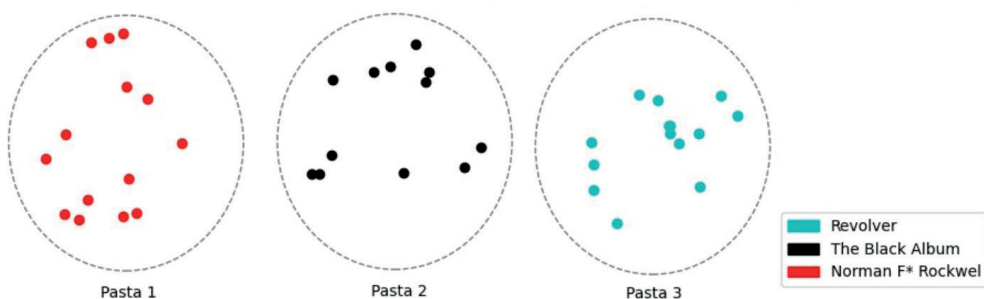



Figura 4.4: Resultado obtido fazendo a separação usando o agrupamento hierárquico nos dados brutos as características do MFCC. A disposição dos pontos no gráfico foram gerados aleatoriamente, apenas para facilitar a visualização.

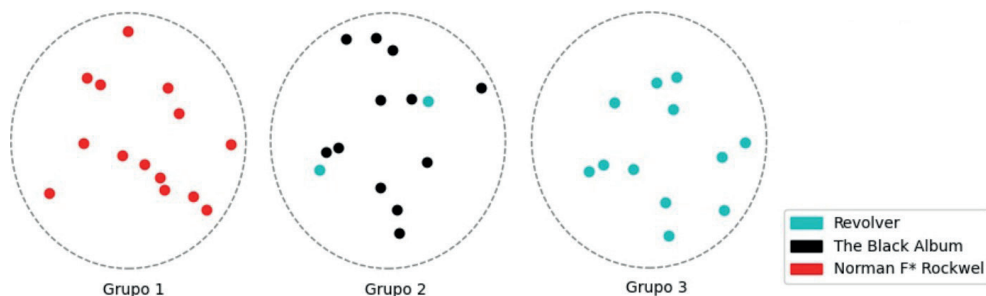



Figura 4.5: Resultado obtido fazendo a separação usando *k*-means nas 3 primeiras componentes principais extraídas das características do MFCC. A disposição dos pontos no gráfico foram gerados aleatoriamente, apenas para facilitar a visualização.

De maneira geral os métodos de agrupamento apresentados nas seçoes anteriores funcionam bem, às vezes cometendo uma quantia pequena de erros como na Figura 4.5 que parecem razoáveis, já que metal e rock são de fato gêneros musicais parecidos.

4.5.2 2º Cenário - Mistura de Gêneros

O 2º cenário é uma aplicação similar a anterior, porém bem mais desafiadora. Os dados consistem em músicas de 8 álbuns aos quais atribuímos 4 gêneros distintos. Foram realizados agrupamentos em $k = 4$ grupos visando tentar resgatar os gêneros originais. Os álbuns usados foram:

- *Revolver - The Beatles* (14 músicas, gênero *Rock*);
- *The Velvet Underground - The Velvet Underground & Nico* (11 músicas, gênero
- *Rock*);
- *Norman F* Rockwell - Lana Del Rey* (14 músicas, gênero *Pop*);
- *Melodrama - Lorde* (11 músicas, gênero *Pop*);
- *Symphonie Fantastique - Hector Berlioz (Compositor)* (5 músicas, gênero *Música Clássica/Soundtrack*);
- *The Lord Of The Rings - The Fellowship Of The Ring (OST) - Howard Shore (Compositor)* (18 músicas, gênero *Música Clássica/Soundtrack*);
- *Kamikaze - Eminem* (13 músicas, gênero *Rap/Hip-Hop*);
- *Born Again - Notorius B.I.G* (15 músicas, gênero *Rap/Hip-Hop*).

Este cenário é consideravelmente mais desafiador de realizar um agrupamento bom. Nenhum método foi capaz de fazer um agrupamento que resgatasse os 4 gêneros atribuídos. O agrupamento hierárquico aglomerativo nos dados brutos foi o que deu o resultado mais interessante (Figura 4.6) sendo capaz de praticamente isolar as músicas de

rock (Pasta 4) e de Clássica/Soundtrack (Pasta 1). Pop e Rap/Hip-Hop ficaram misturados. A Pasta 2 contém principalmente um dos álbuns de Pop e a Pasta 3 mistura os dois álbuns de Rap/Hip-Hop com o outro de Pop.

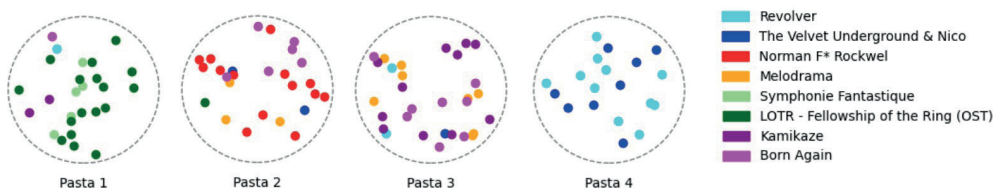



Figura 4.6: Ilustração do resultado para o agrupamento hierárquico com os dados brutos. A disposição dos pontos no gráfico foram gerados aleatoriamente, apenas para facilitar a visualização.

A Figura 4.7 ilustra o agrupamento usando o *k*-means sobre o menor número de componentes principais que explicam 90% da variância. Neste caso podemos ver uma clara anomalia na pasta 3 que acaba pegando uma quantidade muito pequena de músicas tentando isolar o *Symphonie Fantastique*. A pasta 4 mistura os álbuns de Rap/Hip-Hop e o *Melodrama* similarmente à Figura 4.6. A Pasta 2 junta os álbuns de rock com um pouco dos outros gêneros e a Pasta 1 faz uma combinação um tanto problemática já que mistura *Clássica/Soundtrack* com *Pop*.

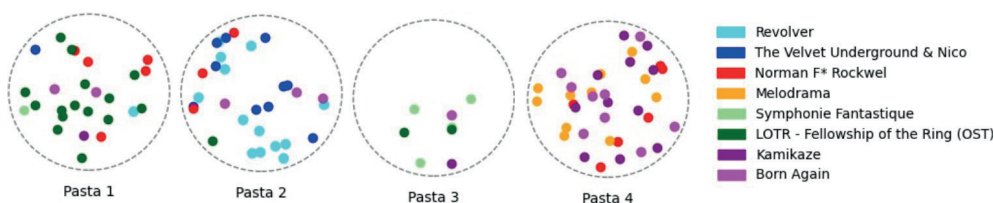



Figura 4.7: Ilustração do resultado para o *k*-means sobre o menor número de componentes principais que explica 90% da variância. A disposição dos pontos no gráfico foram gerados aleatoriamente, apenas para facilitar a visualização.

CONCLUSÃO

Realizar o agrupamento de músicas usando os características do MFCC funciona relativamente bem. Os resultados da sessao anterior mostram que os resultados obtidos no banco de teste são até certo ponto generalizáveis.

O problema, porém, é bastante difícil e há espaço para melhoras. Os agrupamento não são perfeitos (apenas em casos bem específicos) e os melhores métodos para cada caso tendem a variar, não havendo um único método que consistentemente se sobressai. Mas em geral todos são capazes de fazer uma agrupamento minimamente razoável. A

abordagem apresentada pode ser usada, por exemplo, para um pré-processamento de um conjunto de músicas deixando para o usúario apenas o trabalho de validar os grupos e ajustar umas poucas músicas, ao invés de ter que selecionar uma por uma.

REFERÊNCIAS BIBLIOGRÁFICAS

Defferrard, M., Benzi, K., Vandergheynst, P., Bresson, X., 2017. FMA: A dataset for music analysis. arXiv:1612.01840.

Futrelle, J., Downie, J.S., 2003. Interdisciplinary research issues in music information retrieval: ISMIR 2000–2002. Journal of New Music Research 32, 121–131.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Networks 13, 411–430.

Knees, P., Schedl, M., 2016. Music Similarity and Retrieval. Springer-Verlag Berlin Heidelberg.

Logan, B., 2000. Mel Frequency Cepstral Coefficients for music modeling, in: Internati- onal Symposium on Music Information Retrieval, pp. 1–13.

McFee, B., Lostanlen, V., McVicar, M., Metsai, A., Balke, S., Thomé, C., Raffel, C., Lee, D., Zalkow, F., Lee, K., Nieto, O., Mason, J., Ellis, D., Yamamoto, R., Battenberg, E., Morozov, V., Bittner, R., Choi, K., Moore, J., Wei, Z., Seyfarth, S., nullmightybofo, Friesch, P., St¨oter, F.R., nú, D.H., Thassilo, Kim, T., Vollrath, M., Weiss, A., Weiss, A., 2019. librosa/librosa: 0.7.1. doi:10.5281/zenodo.3478579. software.

Mouselimis, L., 2021. ClusterR: Gaussian Mixture Models, K-Means, Mini-Batch- Kmeans, K-Medoids and Affinity Propagation Clustering. URL: https://CRAN. R-project.org/package=ClusterR. R package version 1.2.5.

Murphy, K., 2012. Machine Learning: A Probabilistic Perspective. Adaptive Computa- tion and Machine Learning series, MIT Press.

Murtagh, F., Legendre, P., 2011. Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm. arXiv:1111.6285 .

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Rabiner, L., Rabiner, L., Juang, B., 1993. Fundamentals of Speech Recognition. Prentice-Hall Signal Processing Series: Advanced monographs, PTR Prentice Hall.

Ward, J., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244.

CLASSIFICAÇÃO DE PROJETOS DA CÂMARA DOS DEPUTADOS

Fernanda Kelly Romeiro Silva
Programa de Pós-Graduação em Ciência da Computação - UFG

Franciele Lobo Pallaoro†
Instituto de Matemática e Estatística - UFRGS

Gabriela Soares Rech
Instituto de Matemática e Estatística - UFRGS

Juliana Sena de Souza
Programa de Pós-Graduação em Epidemiologia - UFRGS

Mathias Giordani Titton
Programa de Pós-Graduação em Engenharia Elétrica - UFRGS

Rodrigo Nunes Moni da Silva
Programa de Pós-Graduação em Ciência da Computação - UFRGS

RESUMO: Este estudo tem como objetivo de classificar as proposições da Câmara dos Deputados brasileira em grandes áreas temáticas, baseando-se apenas nas palavras-chaves de cada proposição. A classificação é feita as técnicas de Random Forest, Gradient Boosting Machine e Generalized Linear Model e a avaliação dos modelos é baseada na métrica de acurácia. Dentre os métodos estudados, o Gradient Boosting Machine apresentou os melhores resultados. As técnicas e metodologias empregadas nesse trabalho podem ser usadas em outras esferas do poder público, como a municipal, comprovando a contribuição deste trabalho.
PALAVRAS-CHAVE: Dados políticos, Câmara dos Deputados, Temas, Machine Learning, Classificaçao

5.1 INTRODUÇÃO

Impulsionados pela Lei da Transparência, a União, os estados e os municípios têm feito cada vez mais esforços em disponibilizar informações sobre despesas, receitas e fornecedores dos diversos setores destes órgãos públicos.

Com a Câmara dos Deputados não é diferente, uma vez que é disponibilizado um vasto acervo de dados que podem ser acessados no site dadosabertos.camara.leg.br. Dentre os dados disponibilizados, podemos encontrar os históricos de votações, informações sobre projetos propostos e votados, despesas de cada gabinete, informações específicas de cada deputado e diversas informações relacionadas a eventos, legislaturas e frentes. Os dados

podem ser acessados tanto via *download* de arquivos (*.csv, .json e .xml*), como via API, na qual é possível fazer requisiçoes mais customizadas de dados.

Recentemente, o Centro de Documentação e Informação da Câmara dos Deputados realizou a classificação das proposições em áreas temáticas, tais como: administração pública, ciência, tecnologia e inovação, defesa e segurança, meio ambiente e desenvolvimento sustentável. Diante disso, a disponibilização desses dados classificados motivou a seguinte questionamento: “seria possível classificar as proposições por tema, baseando-se somente nas palavras-chave de cada uma?”

O objetivo deste trabalho é tentar obter respostas à esta questao através da utilização de alguns modelos de machine learning. Neste trabalho propomos a utilização de *Random Forest*, *Gradient Boosting Machine* e *Generalized Linear Model* com regularização LASSO para realizar a classificação dos projetos em temas. Além disso, será feita a avaliação dos modelos utilizando a métrica de acurácia, a fim de verificar a qualidade dos modelos utilizados e assim selecionar o modelo que, mais satisfatoriamente, se encaixe melhor neste contexto. Para a classificação, serao considerados como entrada os dados referentes às proposições da Câmara dos Deputados no período de 1990 a 2019. Para obter e manipular os dados foi usado *Python* e para criação e análise de modelos foi utilizado a linguagem de programação R (R Core Team, 2019), versão 3.6.2. O código de todos os *scripts* e dados utilizados estão disponíveis no GitHub[1].

5.2 DESCRIÇÃO DOS DADOS

Os dados utilizados são relacionados as proposições da Câmara dos Deputados entre os anos de 1990 e 2019. *Proposição*, também chamada de proposta, é basicamente uma denominação genérica para toda matéria submetida à apreciação da Câmara. O processo legislativo compreende a elaboração, análise e votação de vários tipos de propostas: leis ordinárias, medidas provisórias, emendas à Constituição, decretos legislativos e resoluções, entre outras. É importante ressaltar que cada tipo de proposta segue um caminho (tramitação) diferente dentro da Câmara, como por exemplo, quem pode apresentá-la, o tipo de votacão, a quantidade de turnos, entre outros.

As proposições tramitam de acordo com as normas constitucionais e com o Regimento Interno da Câmara (camera.leg.br), no qual as proposições em tramitação na Casa são identificadas por suas siglas, seu número (cada série iniciada em uma legislatura) e o ano em que foi proposta. Entre os principais tipos de propostas, podemos destacar: PEC (Proposta de Emenda à Cons- tituição), PLP (Projeto de Lei Complementar), PL (Projeto de Lei), MPV (Medida Provisória), PLV (Projeto de Lei de Conversão).

O Portal de Dados Abertos da Câmara dos Deputados fornece um vasto conjunto de dados referentes às proposições apresentadas à Câmara dos deputados para deliberação

1 https://github.com/rodrimoni/ProjetoFinalCursoDeVerao

Câmara dos Deputados - Brasil (2020). Elas são disponibilizadas em arquivos (*.csv, .json, .xml*) e via API, divididas por ano. Cada proposição possui as seguintes informações: identificador universal (URI), sigla, número, ano, ementa, palavras-chave, informações sobre a tramitação mais recente, proposições a que se relacionam e entre outras.

O Portal de Dados Abertos da Câmara também disponibiliza informações sobre a classificação temática das proposições, porém em arquivos distintos. Os temas podem ser obtidos via download de arquivos ou por API e também estão separados por anos, bem como os arquivos relacionados às proposições. Cada registro corresponde a uma área temática na qual uma proposição foi classificada pelo Centro de Documentação e Informação da Câmara.

Ao todo, as proposições se dividem em 31 temas específicos. Estes temas estão associados de acordo com a natureza das proposições, sendo elas: administração pública; agricultura, pecúaria, pesca e extrativismo; arte, cultura e religião; cidades e desenvolvimento urbano; ciência, tecnologia e inovação; ciências sociais e humanas; comunicações; defesa e segurança; direito civil e processual civil; direito constitucional; direito e defesa do consumidor; direito e justiça; direito penal e processual penal; direitos humanos e minorias; economia; educação; energia, recursos hídricos e minerais; esporte e lazer; estrutura fundiária; finanças públicas e orçamento; homenagens e datas comemorativas; indústria, comércio e serviços; meio ambiente e desenvolvimento sustentável; política, partidos e eleições; previdência e assistência social; processo legislativo e atuação parlamentar; relações internacionais e comércio exterior; saúde; trabalho e emprego; turismo; viação, transporte e mobilidade. É interessante notar que algumas proposições se enquadram em mais de um tema, e, por consequência, os arquivos trazem múltiplas linhas/entradas com os mesmos identificadores da proposição, uma para cada área temática associada à proposição.

Para a realização deste trabalho, somente as informações de sigla, número, ano, palavras-chave e temas foram utilizadas. Nas Figuras 5.1 e 5.2, podemos verificar a distribuição de propostas por ano e a distribuição de propostas por temas, respectivamente. Podemos notar que os temas mais discutidos na Câmara dos deputados são comunicação, administração pública, trabalho e emprego, e finanças públicas e orçamento. Já em relação à quantidade de proposições por ano, é possível perceber que até o início dos anos 2000 poucas propostas eram apresentadas em cada ano e, em 2003, houve um aumento considerável que foi mantido até o último ano analisado.

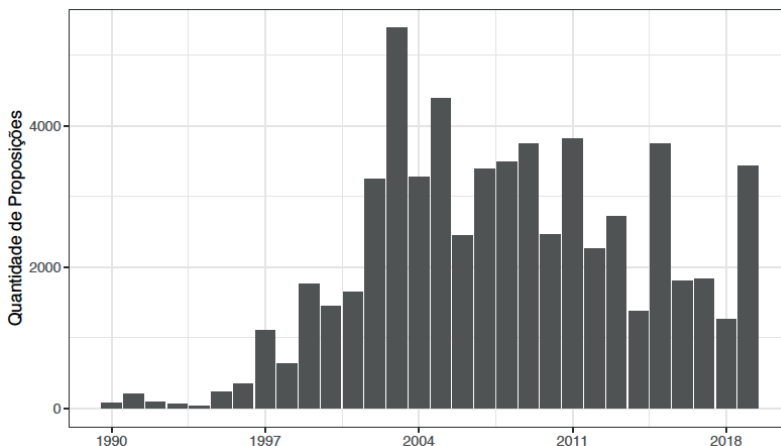



Figura 5.1: Distribuição das propostas por ano.

5.3 PREPARAÇÃO DOS DADOS

Nesta seção serão descritos os procedimentos utilizados para obtenção e pre- paração dos dados, abordando em detalhes os algoritmos e a estrutura de dados utilizados.

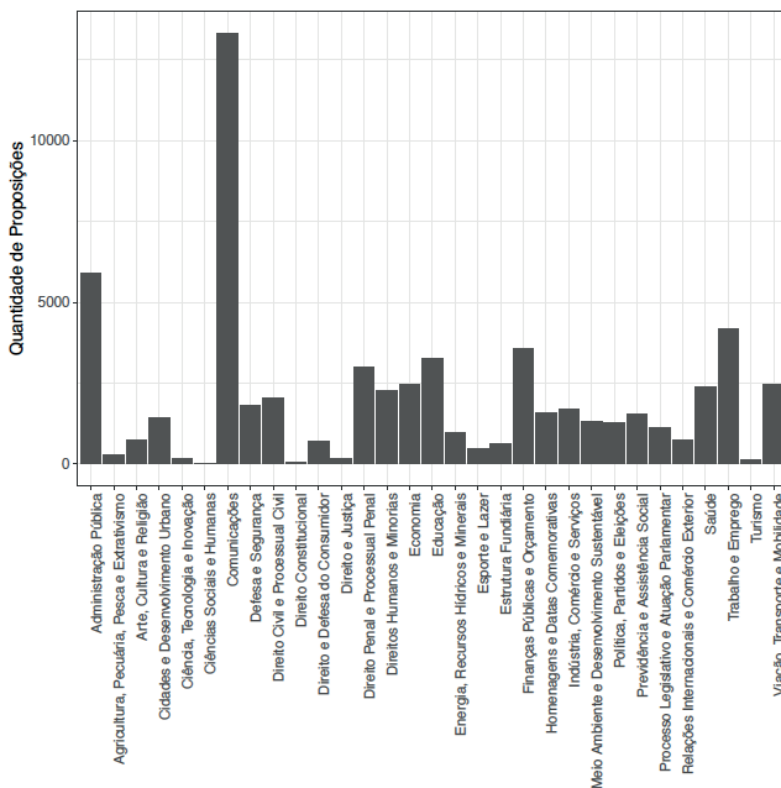



Figura 5.2: Distribuição das propostas por tema.

5.3.1Obtenção dos dados

Os dados foram obtidos através de *download* direto dos dados brutos de proposições e de temas (arquivos em formato *.json*), do site Portal de Dados Abertos da Câmara dos Deputados, disponível na aba Arquivos. Para obter os dados referentes aos anos de nossa análise (1990-2019), foi criado um *script* simples em *Python*, que realizava um *download* dos arquivos e limpava algumas informações desnecessárias (e.g. informação sobre tramitações e *status*) de Proposições e de Temas.

Exemplo de um objeto do conjunto de dados de Proposições:

```
1  {
2      "siglaTipo": "PL",
3      "numero": 402,
4      "ano": 2019,
5      "ementa": "Institui o Programa Cidade Amiga do Idoso.",
6      "descricaoTipo": "Projeto de Lei",
7      "ementaDetalhada": "",
8      "keywords": "Criacao, Programa Cidade Amiga do Idoso,
           qualidade de vida."
9      }
```

Exemplo de um objeto do conjunto de dados de Temas:

```
1  {
2          "uriProposicao": "https://dadosabertos.camara.leg.br/api/v2
               /proposicoes/501638",
3          "siglaTipo": "PL",
4          "numero": 402,
5          "ano": 2019,
6          "codTema": 44,
7          "tema": "Direitos Humanos e Minorias",
8          "relevancia": 0
9  }
```

5.3.2 Preparação dos dados

Após a obtenção dos dados foi necessário prepará-los para seu uso na criação e análise de modelos de aprendizagem estatística. Para isso foi necessário a criação de outro *script* em *Python* que mesclasse o conjunto de Proposições com o conjunto de Temas, de modo que o resultado final tivesse apenas um objeto para cada proposição com as seguintes informações: *index*, tema e palavras-chave. Para não aumentar a complexidade do

problema todas proposições com mais de um tema foram ignoradas, visto que já tínhamos um grande número de categorias (temas). O funcionamento básico do algoritmo é descrito a seguir:

1. Leitura dos objetos nos arquivos de Tema;
2. Criação de um grande dicionário, onde cada índice é indicado pela *sigla + número + ano* de uma proposição e o valor é uma lista de temas;
3. Leitura dos objetos nos arquivos de Proposições;
4. Para cada proposição lida, verificar se o dicionário de temas possuía essa proposição. Esse passo era necessário, pois algumas proposições não possuíam temas.

   a. Caso possuísse: verificar se possuía apenas **um** tema.
      I. Se sim: adicionava no resultado final um objeto contendo apenas *index*, *keywords* e tema;
      II. Caso contrário: ignorava.

   b. Caso contrário: ignorava.

Exemplo de um objeto do conjunto de dados de Proposições mescladas com Temas:

```
1 {
2       "index": "PL4022019",
3       "keywords": "Criacao, Programa Cidade Amiga do Idoso,
            qualidade de vida.",
4       "tema": "Direitos Humanos e Minorias"
5 }
```

Após a mesclagem, foi obtido um conjunto de dados relativamente enxuto de apenas 16MB, com cerca de 60.000 proposições diferentes, uma grande redução de tamanho, visto que inicialmente, sem a limpeza dos dados, o tamanho do conjunto de dados girava em torno de 200MB.

5.4 DESCRIÇÃO DAS TÉCNICAS

Nesta seção serão abordadas as técnicas e metodologias empregadas, os modelos utilizados e as métricas escolhidas.

5.5 PROCESSAMENTO DO TEXTO

Para a criação da matriz de delineamento dos modelos foram utilizadas somente as palavras-chave. A matriz foi criada com as entradas de cada linha correspondendo a cada proposição e, para as colunas, foram empregadas duas metodologias: a primeira usando as palavras-chave completas (ou seja, duas ou mais palavras podem constituir uma coluna, por exemplo: “recibo fiscal”) e o outro método separando todas as palavras em colunas (por exemplo, “recibo” e “fiscal” serão alocados em duas colunas diferentes), esses métodos são chamados de *tokenização*. Assim, as entradas da matriz são uma contagem simples da frequência com que um termo ocorre em todas as palavras-chave de uma proposição (Indurkhya and Damerau, 2010). As Figuras 5.3 e 5.4 apresentam dois esquemas mostrando como foi criada a matriz de delineamento para os modelos.

Como pré-processamento, foram removidas as *stop words*, i.e. palavras que não agregam significado, tais como preposições, artigos e conjunções. Também foi aplicado uma função para remover os termos esporádicos, ou seja, termos que aparecem em menos de 1% das observações. Isso diminui consideravelmente o número de variáveis e, com isso, a complexidade e o poder computacional necessitados para ajustar os modelos de aprendizagem estatística é menor (Indurkhya and Damerau, 2010).

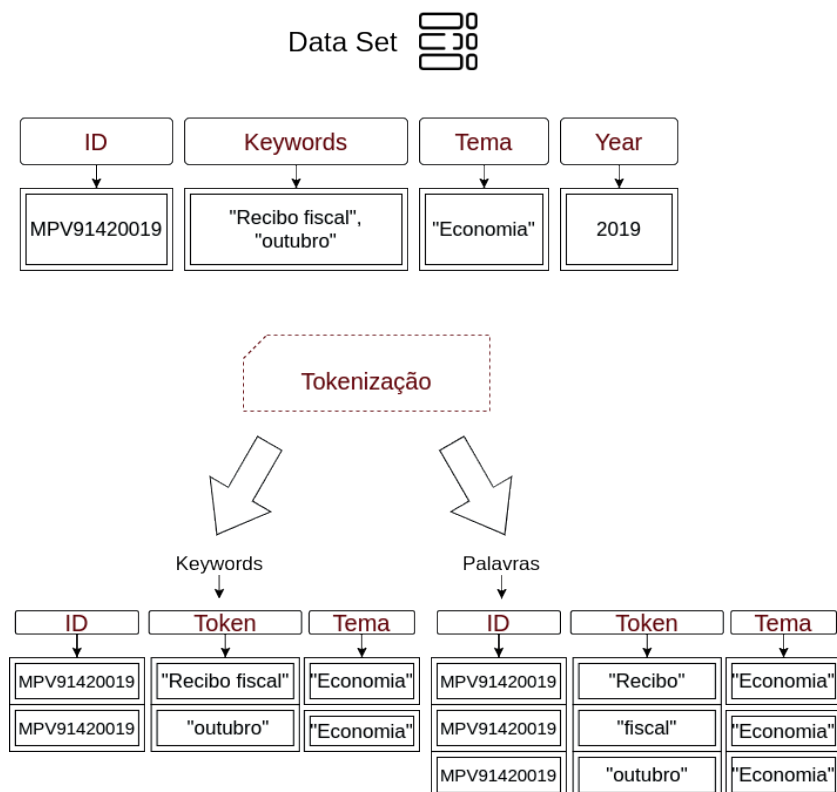



Figura 5.3: Construção da matriz de delineamento dos modelos: *tokenização*

5.5.1 Métodos Empregados

Em seguida, foram implementados os seguintes modelos: *Gradient Boosting Ma- chine* (GBM), *Random Forest* (RF) e o modelo linear generalizado (*Generalized Linear Model* - GLM) multinomial com regularização LASSO.

Gradient Boosting Machine (GBM)

O objetivo do *boosting* é melhorar o desempenho de um único modelo, ajustando muitos modelos e combinando-os para predição. Entretanto, algumas vantagens das árvores de decisão que são sacrificadas pelo *boosting* são a velocidade, a interpretabilidade e, no caso particular do *AdaBoost* , a robustez contra a sobreposição de distribuições de classe e, principalmente, a identificação incorreta dos dados de treinamento. Um modelo GBM é uma generalização do aumento de árvores que tenta suavizar esses problemas, de modo a produzir um procedimento preciso e eficaz (Hastie et al., 2009).

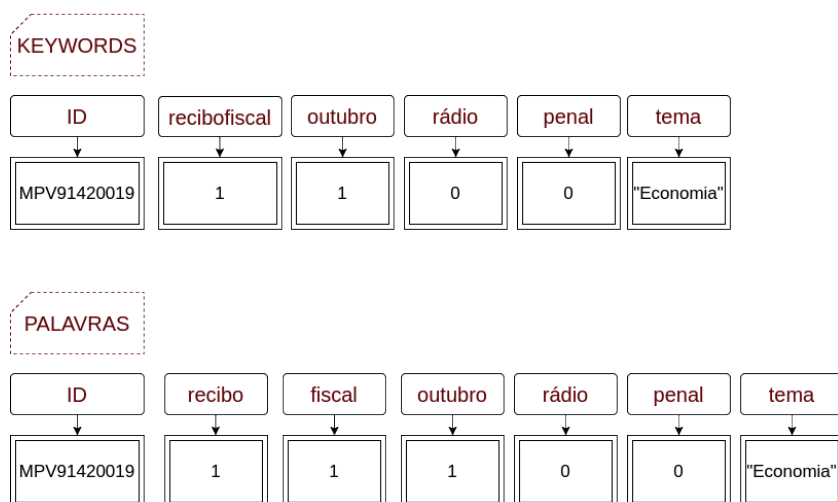



Figura 5.4: Construção da matriz de delineamento dos modelos: contagem da frequência dos *tokens*

Random Forest (RF)

Com as árvores de decisão, tem-se graficamente quais eventos podem acontecer, qual decisão pode ser tomada e quais os resultados associados com combinações de decisões e eventos. Baseado nisso, o método de random Forest é um conjunto das árvores de decisão, em que as previsões individuais das árvores são combinadas para formar uma única previsão (Breiman, 2001).

Generalized Linear Model (GLM)

Os modelos GLM são modelos de regressão em que se assume uma distribuição pertencente a família exponencial para a variável resposta e a resposta média é vinculada às covariáveis por meio de uma função de ligação. A regressão logística multinomial é usada para variáveis resposta com mais de duas categorias e estima a probabilidade condicional da resposta, dado as covariáveis. Neste caso, tipicamente utiliza-se a função de ligação *logit* (McCullagh and Nelder, 1989).

Least Absolute Shrinkage and Selection Operator (LASSO)

O LASSO é um método alternativo de mínimos quadrados generalizados que penaliza a soma dos coeficientes absolutos. O LASSO leva a uma solução esparsa quando o parâmetro de ajuste λ é suficientemente grande. À medida que o valor do parâmetro de ajuste é aumentado, todos os coeficientes são ajustados para zero. Como reduzir os parâmetros para zero os remove do modelo, o LASSO é uma boa ferramenta de seleção de variáveis. Sendo assim, consideramos ainda o modelo linear generalizado multinomial com penalidade LASSO (Hastie et al., 2009).

5.5.2 Métricas

É necessário definir métricas para a escolha do melhor modelo aplicado. A matriz de confusão, por exemplo, é uma tabela que permite a visualização do desempenho do algoritmo, visto que nas linhas são representados os valores verdadeiros, enquanto nas colunas os valores preditos (ou vice-versa) (Hastie et al., 2009). Já a acurácia, trata-se da proporção de resultados corretos que o classificador alcançou, ou seja, a razão entre as predições corretas pelo total (Hastie et al., 2009). Neste trabalho, foram utilizadas como métricas para escolha do melhor modelo a matriz de confusão e a acurácia.

5.6 RESULTADOS E DISCUSSÃO

Os resultados obtidos através dos três modelos estao dispostos abaixo de forma a apresentar primeiramente o GBM, seguido do RF e por último o GLM. Os dados foram separados em conjunto de dados de treinamento (70% do banco), utilizado para treinar o algoritmo e dados de teste (30% do banco) para testar o modelo final. Todos os resultados apresentados aqui foram obtidos utilizando a metodologia para construção da matriz de delineamento baseado em palavras únicas e não nas palavras-chave completas, pois com esta configuração, os modelos ajustados obtiveram uma maior acurácia no banco de teste.

Para a modelagem, os dados finais continham 357 variáveis (termos), e mais a variável resposta (tema) obtidos após a remoção dos termos esporádicos. Caso os

passos descritos para o pré processamento não fossem implementados, isto resultaria num total de 17.769 termos. Caso utilizadas todas essas palavras como variáveis, a matriz associada teria um tamanho maior que a capacidade de memória que pode ser alocada, impossibilitando sua manipulação no R. Os modelos foram ajustados através do pacote *h2o* versão 3.28.0.3 para R. Ressalta-se que este pacote foi escolhido por ser otimizado para lidar melhor com grandes volumes de dados. As análises foram feitas em um computador Linux Mint 19.3 Cinamon, processador AMD-FX(tm)-6300 Six-Core Processor x3 e 8GB de RAM. A execução dos modelos GBM e RF levou cerca de 20 a 30 minutos enquanto que o GLM levou 60 segundos.

Gradient Boosting Machine

O GBM foi o primeiro modelo aplicado. O número de árvores utilizadas foi igual a 200 e a distribuição empregada foi a multinomial. A Figura 5.5 apresenta o erro de classificação no banco de treinamento.

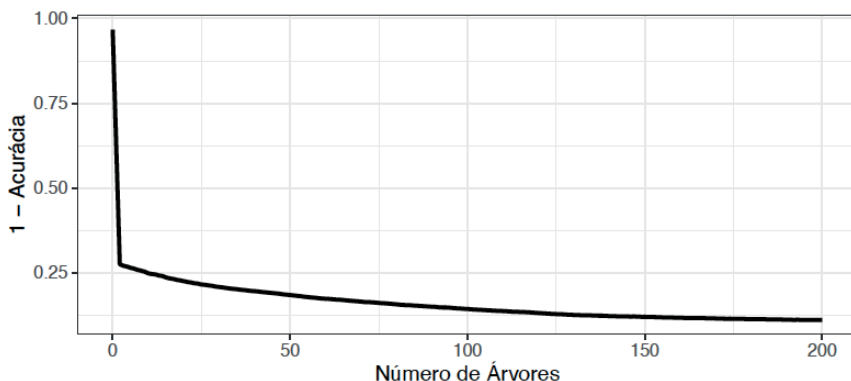



Figura 5.5: Erro de classificação para o *Gradient Boosting Machine* em função do número de árvores.

A métrica apresentada mostra um declínio conforme o número de árvores vai aumentando, indicando que para minimizar o erro de classificação, necessita-se de uma grande quantidade de árvores. Porém, quando este número é superior à 100, a diminuição do erro não varia muito. A acurácia no banco de treinamento foi de 88,98%. Dentre as 357 variáveis, as 15 variáveis principais estão representadas na Figura 5.6, cuja soma de importância equivale à 43,36%. A variável “rádio” foi a mais importante (17,17%), seguida por “penal” (4,54%).

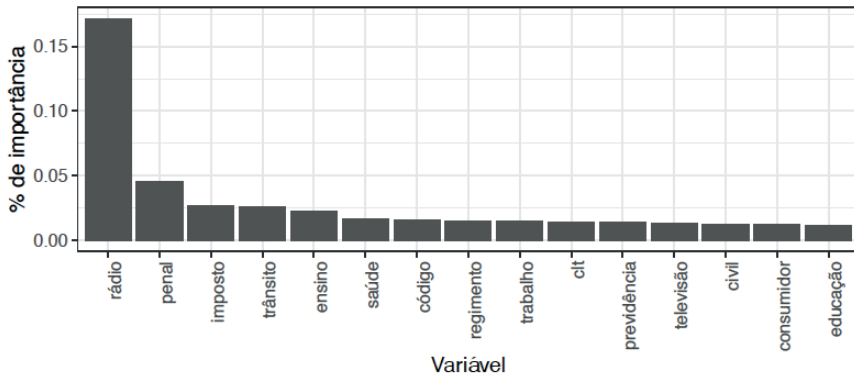



Figura 5.6: As 15 variáveis mais importantes de acordo com o método GBM.

No caso da primeira, foi a palavra que mais apareceu em “comunicação”, com frequência de 11.768. Além de ser o tema com maior número de proposições, a categoria obteve uma acurácia de 99,06%. A Figura 5.7 apresenta a nuvem de palavras das categorias onde as duas variáveis mais importantes se encontram em maior frequência. Já no caso de “penal”, evidentemente é a palavra que mais aparece em ‘direito penal e processual penal’, em 3.613 vezes. Esta categoria obteve acurácia de 94,85%.

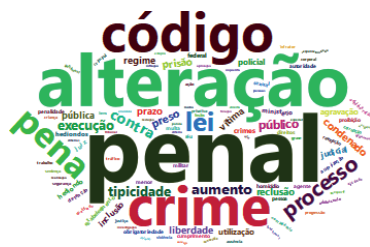
(a) Comunicação

(b) Direito penal e processual penal

Figura 5.7: Nuvem de palavras para os temas referente as variáveis mais importantes do GBM.

Em seguida, ao aplicar o modelo treinado no banco de teste, obteve-se uma acurácia de 77,56%. Ao se analisar as categorias, nota-se que “ciência, tecnologia e inovação”, “direito constitucional” e “direito e justiça” não classificaram bem: para o primeiro tema a acurácia foi de 6,25% (apenas 3 proposições classificadas corretamente de um total de 48), para o segundo foi de 9,09% (1 de 11) e para o último, 1,79% (1 de 56). Em contra partida, olhando as duas categorias que dizem respeito às variáveis mais importantes da Figura 5.6, para o tema “comunicação” o algoritmo classificou incorretamente apenas 85 de 4.073 (97,91% de acurácia) proposições e “direito penal e processual penal”, 100 de 906 (88,96% de acurácia). A matriz de confusão para o modelo GBM está representada na Tabela 5.2, que encontra-se no Apêndice.

Random Forest

O segundo modelo empregado foi o *Random Forest*. O número de árvores utilizadas também foi de 200. Não é nenhuma surpresa que o comportamento do erro de classificação seja similar ao encontrado pelo GBM, visto que esta medida tende a ser cada vez menor conforme a complexidade do modelo (número de árvores) aumenta. Observa-se que, a partir de 76 Árvores, o erro estabiliza em torno de 0,22, alterando somente a partir da terceira casa decimal, em que, neste caso, foi de 0,2299, enquanto que com 200 árvores observou-se um erro de 0,2264. As principais variáveis mais importantes pelo RF estão apresentadas na Figura 5.9.

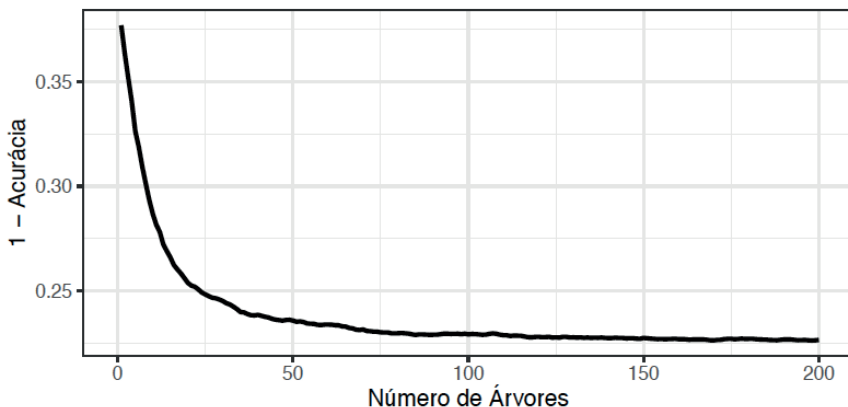



Figura 5.8: Erro de classificação para o *Random Forest* em função do número de árvores.

As variáveis apresentadas na Figura 5.9 representam 30,6% da importância total através do *Random Forest*. Novamente "rádio" foi a mais importante (9,81%), mas dessa vez foi seguida de "aprovação" (3,81%). A variável "penal" foi a quarta mais importante, representando 2,29% do percentual de importância total. A Figura Figure 5.7 mostra que no tema de 'Comunicação' a variável "aprovação" apareceu um número significativo de vezes, sendo a segunda palavra com maior número de aparições, 7.988. Isso mostra que, novamente, o tema com maior quantidade de proposições é o de maior relevância para o modelo. A acurácia no banco de treinamento foi de 77,36%, sendo que "comunicações", "educação", "direito penal e processual penal" e "homenagens e datas comemorativas" tiveram acurácia superior à 90% (97,83%, 92,88%, 91,19% e 90,69%, respectivamente). Já os piores classificados foram: "ciências sociais e humanas", cuja acurácia foi de 0% (errou a classificação da única proposição desse tema); "ciência, tecnologia e inovação", com 2,94% (3 acertos de um total de 102); "direito e justiça", com acurácia de 3,06% e "turismo", que errou 77 de um total de 83 proposições, ficando assim com uma porcentagem de acertos de apenas 7,23%. Não é nenhuma surpresa que a palavra que mais aparece dentro do tema de "turismo" seja a própria palavra, conforme Figura 5.10.

Figura 5.9: As 15 variáveis mais importantes do *Random Forest.*

Em relação ao conjunto de teste, uma vez que foi aplicado o modelo de RF nos dados, obteve-se uma acurácia de 77,7%, um pouco melhor do que o obtido no banco de treino. Entre os temas, novamente "comunicações" foi o melhor classificado, com erro de classificação de somente 1,87%. Os temas que obtiveram também bons resultados foram "direito penal e processual penal", "educação", "homenagens e datas comemorativas" e "relações internacionais e comércio exterior", com acurácia de 92,6%, 92,35%, 91,89% e 90,41%, respectivamente. É curioso observar como "relações internacionais e comércio exterior" conseguiu um ótimo resultado (foram apenas 21 erros, de um total de 219).

Figura 5.10: Nuvem de palavras para o tema de "turismo".

É de se esperar que as categorias com menor número de proposições não sejam bem classificadas e de fato é o que acontece. Dentre os seis temas com piores acurácias, cinco possuíam menos de 100 proposições, são elas: "direito e justiça", "ciência, tecnologia e inovação", "direito constitucional", "turismo" e "agricultura, pecuária, pesca e extrativismo", cuja acurácia foi de 1,79%, 2,08%, 9,09%, 15,63% e 18,99%. A Tabela 5.3 fornece a matriz de confusão para o banco de teste aplicado ao modelo do *Random Forest*. Novamente os temas estão codificados devido ao tamanho da matriz, que possui 31 categorias.

Modelo Linear Generalizado

Por último, foi ajustada uma regressão logística multinomial aos dados. O modelo convergiu com 50 iterações. Como foi usado penalização LASSO foi, feita uma validação cruzada (*5-fold*) para encontrar o melhor valor de λ, que foi aproximadamente 0,0004, muito próximo de zero. Este resultado é evidência de que que não fez diferença usar essa penalidade, já que conforme o λ se aproxima de zero, o LASSO se aproxima do estimador de mínimos quadráticos ordinários.

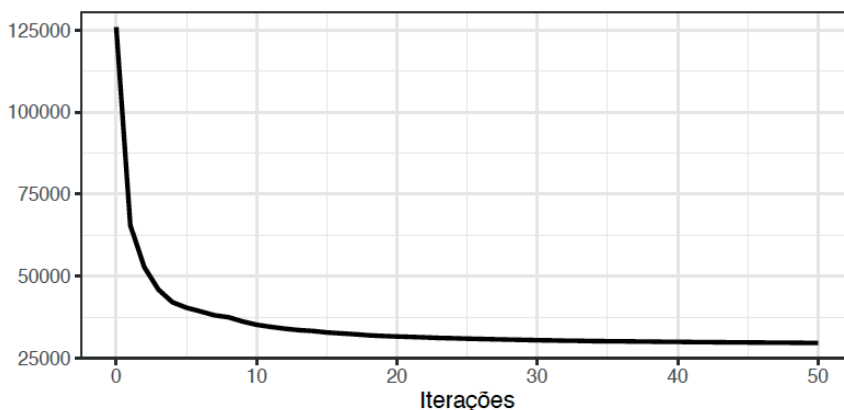



Figura 5.11: Gráfico da –log-verossimilhança em função do número de iterações no Modelo Linear Generalizado.

Para encontrar a solução do modelo deve-se minimizar o negativo do logaritmo da função de verossimilhança, ou seja, maximizar a verossimilhança. Como visto na Figura 5.11 após 10 iterações o valor estabilizou e na iteração 50 obteve o valor de 29.665,65, atingindo o mínimo e assim o algoritmo usado para otimização da função convergiu.

Na Figura 5.12 se encontram as variáveis que tiveram os maiores valores dos coeficientes padronizados no modelo. Palavras como “rádio”, “penal” e “aprovação” aparecem entre aquelas de maior magnitude no GLM, assim como apareceram no GBM e RF. Este comportamento já era esperado, visto o que já foi discutido anteriormente.

Fora isto, algumas palavras com os coeficientes mais elevados foram um pouco diferentes das variáveis mais importantes dos outros dois métodos. Aqui o termo “código” teve o maior magnitude de coeficiente associado de 10,282, seguido de “penal” que também apareceu no RF e GBM, e “eleitoral”, com 7,988 de magnitude padronizada. A *wordcloud* da Figura 5.13 mostra que a variável “eleitoral” foi a mais encontrada no tema “política, partidos e eleições”. Novamente, a relação entre a palavra com maior frequência dentro de um tema tem relação direta com o próprio tema, com “eleitoral” possuindo 1.423 citações entre as proposições de “política, partidos e eleições".

Figura 5.12: As 15 variáveis com magnitude dos Coeficientes Padronizados no Modelo Linear Generalizado

Figura 5.13: Nuvem de palavras para o tema “política, partidos e eleições”

Com relação aos dados de treinamento, o modelo GLM obteve uma acurácia de 79,66%, inferior ao encontrado pelo GBM, mas superior ao do RF. Quanto à quantidade de acertos obtidos no conjunto de dados de teste, obteve-se 74,93% de acertos, resultado inferior aos outros dois métodos.

A taxa de acertos para a categoria “política, partidos e eleições” no banco de teste foi de 79,95%, uma das maiores acurácias entre as categorias, ficando atrás somente de “comunicações”, “direito penal e processual penal”, “homenagens e datas comemorativas”, “educação” e “trabalho e emprego”.

5.7 CONCLUSÃO

Como é possível observar nos resultados obtidos, a acurácia do modelo final escolhido, o GBM, foi relativamente alta, sendo 88,98% no banco de treinamento e 77,56% no banco de teste. Através da análise do GBM, foi possível perceber que as palavras mais importantes foram justamente aquelas relacionadas aos temas que tiveram uma melhor predição e que tinham um maior número de proposições associadas a eles. Por exemplo: a palavra "rádio" foi considerada a palavra mais importante, e ela aparece, predominantemente, nas proposições do tema " comunicações", assim como o contrário também acontece: os temas que apresentaram baixa acurácia em sua predição eram os que possuíam um menor número de proposições associadas. Portanto, se aumentássemos o número de proposições dos temas que foram mal classificados, possivelmente obteríamos melhores resultados na predição.

Por fim, atingimos o objetivo do trabalho de classificar as proposições apresentadas na Câmara dos Deputados por temas, através das três metodologias propostas. Portanto, este estudo é de grande contribuição, visto que as técnicas e metodologias aplicadas nesse trabalho podem ser usadas em outras esferas do poder público como a municipal, para a classificação temática das proposições da Câmara dos Vereadores, por exemplo.

Para trabalhos futuros, existe um vasto campo ainda não explorado. Primeiramente, melhorar o ajuste dos modelos empregados para atingir uma maior acurácia utilizando a combinação do texto completo de cada ementa com as *keywords*. Além disso, realizar a classificação de outras proposições da Câmara dos Deputados que hoje encontram-se sem temas e avaliar o resultado obtido. Assim como, testar as metodologias e técnicas utilizadas nesse trabalho em outros contextos, como: Senado Federal e Câmaras Estaduais e Municipais.

5.8 APÊNDICE

Nesta seção são apresentadas as tabelas com a codificação numérica de temas empregados na análise bem como as matrizes de confusão referente ao teste de cada uma das técnicas utilizadas.

Código	Temas
1	Administração Pública
2	Agricultura, Pecúaria, Pesca e Extrativismo
3	Arte, Cultura e Religião
4	Cidades e Desenvolvimento Urbano
5	Ciência, Tecnologia e Inovação
6	Ciências Sociais e Humanas
7	Comunicações
8	Defesa e Segurança
9	Direito Civil e Processual Civil
10	Direito Constitucional
11	Direito Penal e Processual Penal
12	Direito e Defesa do Consumidor
13	Direito e Justiça
14	Direitos Humanos e Minorias
15	Economia
16	Educação
17	Energia, Recursos Hídricos e Minerais
18	Esporte e Lazer
19	Estrutura Fundiária
20	Finanças Públicas e Orçamento
21	Homenagens e Datas Comemorativas
22	Indústria, Comércio e Serviços
23	Meio Ambiente e Desenvolvimento Sustentável
24	Política, Partidos e Eleições
25	Previdência e Assistência Social
26	Processo Legislativo e Atuação Parlamentar
27	Relações Internacionais e Comércio Exterior
28	Saúde
29	Trabalho e Emprego
30	Turismo
31	Viação, Transporte e Mobilidade

Tabela 5.1: Codificação dos temas de acordo com a numeração empregada

Tabela 5.2: Matriz de Confusão para o banco de teste do método GBM

Tema	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1.268	1	10	6	4	0	15	26	13	0	14	2	4	16	59	19	15
2	7	16	0	0	0	0	0	1	1	0	0	0	0	0	8	1	3
3	34	1	59	1	0	0	2	2	4	1	1	0	0	4	3	10	2
4	28	2	1	218	1	0	0	9	4	0	0	2	0	2	16	3	8
5	6	1	1	0	3	0	4	0	0	0	2	1	0	0	6	2	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	23	0	0	0	0	0	3.988	4	4	0	0	5	0	6	5	1	1
8	58	0	3	4	0	0	11	343	4	0	26	0	0	15	10	4	2
9	41	0	2	7	0	0	5	8	440	0	5	4	3	9	12	2	0
10	6	0	0	0	0	0	0	0	1	1	1	0	0	0	1	0	0
11	16	0	0	0	0	0	1	22	7	0	806	2	1	12	4	3	1
12	11	0	0	0	0	0	9	0	4	0	0	121	0	0	6	1	1
13	24	0	0	1	0	0	2	2	15	0	4	0	1	0	0	0	0
14	39	2	5	3	1	0	3	7	8	2	14	1	0	443	10	13	3
15	99	4	0	7	0	0	3	9	7	0	5	4	0	6	444	10	12
16	24	2	2	1	0	0	6	4	0	0	1	0	0	8	7	892	0
17	25	1	1	0	0	0	1	1	1	0	1	1	0	0	18	1	171
18	29	1	3	0	0	0	3	7	1	0	1	0	0	6	5	14	0
19	24	11	1	1	2	0	0	2	3	0	0	0	1	2	15	2	5
20	58	4	1	3	1	0	5	5	4	0	2	2	2	6	51	9	7
21	12	0	16	0	0	0	0	0	0	0	0	0	0	2	0	3	0
22	38	4	0	4	2	0	15	7	13	0	3	53	0	4	22	2	12
23	42	3	1	2	0	0	0	2	2	0	11	0	0	1	8	6	10
24	48	1	0	0	0	0	0	1	1	0	3	0	0	1	0	0	0
25	16	0	0	0	0	0	1	1	1	0	2	0	0	14	8	3	0
26	43	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	0
27	12	0	1	0	0	0	0	2	1	0	0	0	0	2	4	2	0
28	30	1	3	1	1	0	5	2	5	0	4	0	0	28	6	19	3
29	48	1	2	6	1	0	3	7	8	0	4	3	3	15	12	18	3
30	6	0	1	1	0	0	4	0	0	0	0	0	0	1	1	2	0
31	37	0	6	77	0	0	2	10	2	0	1	1	1	5	8	7	5
Total	2.152	56	120	343	16	0	4.088	485	554	4	911	203	16	609	750	1.050	264

18	19	20	21	22	23	24	25	26	27	28	29	30	31	Erro	Taxa
3	7	32	11	13	6	74	20	26	1	29	41	1	23	0,28	491/1.759
0	23	4	0	4	1	0	1	0	0	3	4	1	1	0,80	63/79
2	0	3	72	2	2	0	1	2	2	3	5	0	12	0,74	171/230
0	6	5	0	6	4	1	3	1	0	4	11	0	110	0,51	227/445
1	0	3	0	6	4	0	0	0	1	5	2	0	0	0,94	45/48
0	0	0	0	0	0	0	0	0	0	0	0	0	0		0/0
0	0	2	0	15	1	0	0	0	2	5	4	1	6	0,02	85/4.073
0	2	7	3	13	6	5	1	1	2	9	5	0	11	0,37	202/545
1	1	9	1	6	3	1	3	0	0	6	11	0	1	0,24	141/581
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0,91	10/11
0	0	2	0	5	4	1	3	2	3	2	7	0	2	0,11	100/906
0	0	0	0	41	0	0	0	0	0	4	0	1	10	0,42	88/209
1	0	0	0	0	0	0	0	2	3	0	1	0	0	0,98	55/56
0	2	2	9	7	2	3	16	6	4	34	10	1	8	0,33	215/658
1	4	44	3	12	4	2	10	2	3	4	18	1	10	0,39	284/728
1	0	6	7	1	2	1	1	0	2	12	11	0	3	0,10	102/994
0	3	4	2	10	8	0	1	1	2	3	4	0	8	0,36	97/268
33	0	4	4	5	1	1	0	0	0	8	7	0	2	0,76	102/135
1	104	7	2	3	7	0	1	0	1	4	2	0	0	0,48	97/201
1	6	837	1	7	6	1	8	3	3	3	6	0	10	0,20	215/1.052
0	0	0	411	0	2	0	0	0	1	1	2	0	6	0,10	45/456
1	5	17	1	265	5	1	1	1	2	26	7	2	11	0,49	259/524
1	6	0	2	21	239	0	0	2	1	15	2	1	10	0,38	149/388
1	0	0	1	0	1	332	0	2	0	1	0	0	1	0,16	62/394
2	0	8	0	1	2	1	355	0	1	7	19	1	2	0,20	90/445
1	0	2	2	1	2	4	1	268	4	1	0	0	1	0,20	67/335
0	1	1	1	1	1	0	0	3	184	1	0	0	2	0,16	35/219
0	3	5	6	40	2	0	6	1	1	524	15	0	5	0,27	195/719
3	3	17	8	6	4	0	16	0	0	16	1035	0	10	0,17	215/1.250
1	0	1	2	0	0	0	0	0	0	0	2	9	1	0,72	23/32
1	0	7	16	12	3	0	2	1	0	8	6	1	532	0,29	219/751
56	176	1.029	565	503	322	428	450	324	223	738	1.237	20	799	0,22	4.149/18.491

Tabela 5.3: Matriz de Confusão para o banco de teste do método RF

Tema	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1.272	2	7	3	0	0	32	29	15	0	17	0	0	9	43	24	24
2	6	15	0	0	0	0	0	1	1	0	0	0	0	0	8	1	0
3	19	0	41	0	0	0	16	1	4	1	1	0	0	3	4	16	1
4	24	1	0	176	0	0	7	9	8	0	0	2	0	8	16	4	4
5	7	1	0	0	1	0	4	1	0	0	4	0	0	0	6	1	2
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	18	0	1	0	0	0	3.997	6	3	0	2	3	1	5	3	1	0
8	46	0	0	2	0	0	18	345	3	0	37	1	0	13	9	5	2
9	41	0	0	4	0	0	9	7	438	0	10	2	2	7	6	1	0
10	5	0	0	0	0	0	0	0	0	1	2	0	0	0	2	1	0
11	16	0	0	0	0	0	2	9	5	0	839	1	0	10	4	2	0
12	11	0	0	0	0	0	12	0	3	0	0	114	0	0	9	1	1
13	27	0	0	0	0	0	4	1	16	1	2	0	1	0	0	0	0
14	47	1	1	1	0	0	8	6	10	1	16	1	1	416	14	20	1
15	110	1	0	4	0	0	15	10	5	0	5	4	0	6	430	14	15
16	19	1	0	0	0	0	9	1	0	0	0	0	0	3	4	918	1
17	31	1	0	1	0	0	3	2	4	0	2	0	0	0	10	2	159
18	37	0	2	0	0	0	4	2	1	0	3	0	0	6	3	13	0
19	21	5	0	1	0	0	4	1	1	0	1	0	0	2	20	1	4
20	47	0	0	0	0	0	13	5	4	0	4	3	0	2	53	9	7
21	7	0	13	0	0	0	7	0	0	0	0	0	0	2	0	3	0
22	37	2	0	1	0	0	21	5	12	0	5	55	0	5	21	23	9
23	38	3	1	0	0	0	15	2	3	0	20	0	0	1	8	3	5
24	33	0	0	0	0	0	2	0	0	0	2	0	0	2	0	1	1
25	24	0	0	0	0	0	2	1	2	0	0	0	0	10	11	5	1
26	31	0	0	0	0	0	3	1	0	0	1	0	0	0	1	2	0
27	5	0	1	0	0	0	5	2	1	0	0	0	0	1	1	0	0
28	26	1	1	0	0	0	18	2	6	0	2	1	0	19	3	27	4
29	39	0	1	1	0	0	8	8	3	0	4	1	2	13	10	18	1
30	4	0	0	0	0	0	1	1	0	0	0	0	0	1	2	2	0
31	34	0	1	43	0	0	6	9	6	0	2	0	0	5	10	10	4
Total	2.082	34	70	237	1	0	4.245	467	554	4	981	188	7	549	711	1128	246

18	19	20	21	22	23	24	25	26	27	28	29	30	31	Erro	Taxa
0	4	33	14	12	10	85	15	23	4	22	40	0	20	0,28	487/1.759
0	27	5	0	5	2	0	1	0	0	0	5	0	2	0,81	64/79
1	0	6	97	1	1	0	1	0	2	3	5	0	6	0,82	189/230
0	6	4	1	4	6	1	3	3	0	5	12	0	141	0,60	269/445
0	1	3	0	5	3	0	0	1	1	5	2	0	0	0,98	47/48
0	0	0	0	0	0	0	0	0	0	0	0	0	0		0/0
0	1	1	0	10	2	1	0	1	1	4	4	0	8	0,02	76/4.073
0	1	4	4	7	2	3	2	1	5	14	6	0	15	0,37	200/545
0	2	14	2	10	1	4	1	0	0	4	14	0	2	0,25	143/581
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,91	10/11
0	0	2	0	3	2	1	0	0	3	2	3	0	2	0,07	67/906
0	0	0	1	41	0	0	1	0	0	3	0	0	12	0,45	95/209
0	0	1	0	0	0	2	0	0	1	0	0	0	0	0,98	55/56
0	2	9	12	7	1	3	11	8	3	37	13	0	8	0,37	242/658
0	3	44	2	6	2	2	7	1	1	5	24	0	12	0,41	298/728
0	0	5	7	2	1	1	3	0	2	7	8	0	2	0,08	76/994
0	2	4	3	11	10	0	1	1	1	5	4	0	11	0,41	109/268
28	0	5	4	5	0	1	2	0	1	5	13	0	0	0,79	107/135
0	102	12	2	4	5	0	1	0	1	4	8	0	1	0,49	99/201
0	2	858	2	5	4	1	9	3	2	5	6	2	6	0,18	194/1.052
0	0	0	419	0	1	0	0	0	0	0	2	0	2	0,08	37/456
0	2	27	2	230	5	0	2	0	2	25	16	0	17	0,56	294/524
0	5	0	3	15	235	0	0	1	1	14	4	0	11	0,39	153/388
0	0	0	1	0	0	348	0	1	0	0	0	0	3	0,12	46/394
0	0	9	0	1	0	0	356	0	1	4	18	0	0	0,20	89/445
0	0	2	2	0	1	6	1	277	3	1	2	0	1	0,17	58/335
0	0	1	1	0	0	0	0	3	198	0	0	0	0	0,10	21/219
0	0	7	8	40	2	0	1	3	1	526	10	0	11	0,27	193/719
0	3	17	10	3	3	4	18	1	1	14	1.061	0	6	0,15	189/1.250
0	0	2	4	1	2	0	0	1	1	0	3	5	2	0,84	27/32
0	0	8	28	4	3	0	1	0	0	7	9	0	561	0,25	190/751
29	163	1.083	629	432	304	463	437	329	236	721	1.292	7	862	0,22	4.124/18.491

Tabela 5.4: Matriz de Confusão para o banco de teste do método GLM

Tema	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1.184	8	11	10	7	0	23	34	27	0	9	0	5	25	70	14	30
2	3	19	2	0	0	0	0	2	1	0	0	0	0	0	8	6	0
3	19	5	57	3	1	0	3	1	3	0	1	0	0	6	3	8	2
4	18	0	0	221	0	0	2	7	8	0	0	1	0	7	18	3	14
5	8	1	0	0	6	0	3	0	0	0	2	1	0	0	3	2	1
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	21	2	1	0	2	0	3.971	7	3	0	1	7	0	4	3	1	2
8	49	1	1	6	0	0	14	329	6	1	26	2	0	20	9	6	2
9	33	1	3	6	0	0	8	6	433	0	6	3	4	9	12	1	1
10	3	0	0	1	0	0	0	1	0	1	1	0	1	2	0	0	0
11	18	0	1	0	0	0	4	28	7	0	797	0	2	12	4	3	0
12	9	0	0	0	0	0	11	0	2	0	0	109	0	0	5	1	1
13	16	0	0	0	0	0	2	2	20	1	2	0	5	1	0	0	0
14	31	1	8	6	1	0	4	11	14	1	12	3	1	415	5	15	4
15	97	6	5	6	0	0	18	7	8	0	5	2	0	4	393	13	23
16	25	2	1	0	1	0	6	1	2	0	3	0	0	13	8	858	3
17	22	3	3	7	1	0	0	1	3	0	0	0	0	1	15	3	161
18	14	0	4	2	1	0	1	12	6	0	1	0	0	8	7	10	1
19	21	15	3	2	5	0	2	1	4	1	0	0	0	1	17	4	8
20	39	3	2	2	1	0	14	6	8	0	2	0	1	12	74	11	11
21	7	0	30	0	0	0	2	1	0	0	0	0	1	1	1	3	0
22	23	6	0	3	6	0	17	9	12	0	2	51	0	5	15	7	8
23	28	4	2	3	1	0	5	1	8	0	14	0	0	1	5	5	12
24	63	0	0	1	0	0	1	1	0	0	1	0	0	5	0	0	1
25	23	2	1	0	0	0	0	1	4	0	0	0	1	19	7	4	1
26	48	0	0	0	0	0	2	0	1	0	1	0	0	0	2	1	0
27	6	0	0	0	1	0	2	6	2	0	2	0	1	1	5	0	0
28	16	1	2	3	3	0	16	6	11	0	2	1	0	23	8	20	6
29	39	2	2	4	1	0	7	8	7	0	4	2	1	13	13	14	2
30	2	0	2	0	0	0	2	2	0	0	0	1	0	1	2	1	1
31	29	0	5	101	2	0	6	10	7	1	5	2	0	8	15	5	9
Total	1.914	82	146	387	42	0	4.146	501	607	6	899	185	23	617	727	1.019	304

18	19	20	21	22	23	24	25	26	27	28	29	30	31	Erro	Taxa
10	8	34	12	19	11	59	22	32	5	23	44	2	21	0,33	575/1.759
0	18	3	2	4	4	0	1	0	0	2	2	0	2	0,76	60/79
5	0	3	75	2	2	1	2	1	3	4	3	1	16	0,75	173/230
4	4	3	0	5	8	1	0	0	0	3	8	1	109	0,50	224/445
0	2	4	0	5	3	0	0	0	1	2	1	1	2	0,88	42/48
0	0	0	0	0	0	0	0	0	0	0	0	0	0		0/0
1	0	2	0	23	2	1	1	1	1	6	1	4	5	0,03	102/4.073
5	1	8	3	10	5	1	0	1	4	10	10	1	12	0,40	216/545
2	3	6	2	9	2	0	2	3	2	6	15	0	3	0,25	148/581
0	0	1	0	0	0	0	0			0	0	0	0	0,91	10/11
3	0	1	0	9	6	3	0	0	2	3	2	0	1	0,12	109/906
1	0	1	0	52	0	0	0	0	0	5	2	1	9	0,48	100/209
1	0	1	0	0	0	1	0	2	2	1	1	0	0	0,91	51/56
3	4	8	11	8	3	1	15	4	2	44	11	1	11	0,37	243/658
7	2	42	3	26	6	0	15	0	2	8	18	1	11	0,46	335/728
4	1	8	5	4	3	1	3	1	2	13	20	0	6	0,14	136/994
3	2	7	2	11	7	0	0	1	0	3	8	0	4	0,40	107/268
34	0	4	5	3	1	0	2	1	0	10	7	0	1	0,75	101/135
2	90	5	0	4	3	0	2	0	3	5	3	0	0	0,55	111/201
3	5	800	1	12	5	2	8	4	2	4	7	2	11	0,24	252/1.052
1	0	0	399	1	1	0	0	0	0	1	2	0	5	0,12	57/456
4	7	14	2	250	9	0	3	2	3	36	11	1	18	0,52	274/524
1	9	1	4	20	235	0	2	2	0	17	2	1	5	0,39	153/388
0	0	0	0	1	2	315	0	1	0	0	0	0	2	0,20	79/394
1	1	13	0	1	0	0	336	1	0	3	24	0	2	0,24	109/445
0	1	2	2	1	2	5	1	260	2	3	0	0	1	0,22	75/335
0	1	2	2	0	1	1	0	6	174	4	0	1	1	0,21	45/219
5	4	7	7	34	8	0	11	1	0	505	13	1	5	0,30	214/719
9	5	15	10	7	2	1	26	2	0	18	1.028	0	8	0,18	222/1.250
1	0	1	4	1	0	0	0	1	0	1	1	5	3	0,84	27/32
2	0	9	16	12	6	0	6	0	1	10	17	1	466	0,38	285/751
112	168	1.005	567	534	337	393	458	325	211	750	1.261	25	740	0,25	4.635/18.491

REFERÊNCIAS BIBLIOGRÁFICAS

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Câmara dos Deputados - Brasil, 2020. Dados abertos da câmara dos deputados. https://dadosabertos.camara.leg.br/. (acessado em 28 de Janeiro de 2020).

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Indurkhya, N., Damerau, F.J., 2010. Handbook of natural language processing. 2 ed., CRC Press.

McCullagh, P., Nelder, J.A., 1989. Generalized linear models. 2 ed., Chapman and Hall/CRC.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https: //www.R-project.org/.

GUILHERME PUMI - É mestre em estatística pela University of California at Davis (USA), mestre e doutor em matemática pela Universidade Federal do Rio Grande do Sul (UFRGS) e possui pós-doutorado em matemática pela UFRGS. É professor do Departamento de Estatística da UFRGS onde também atua no Programa de Pós-Graduação em Estatística como professor permanente tendo já sido coordenador do programa. Seus interesses de pesquisa são focados principalmente no estudo de séries temporais, em especial no desenvolvimento de novos modelos para séries não-Gaussianas, modelos GARMA, estimação paramétrica e semi-paramétrica, teoria assintótica, cópulas em séries temporais, séries temporais com longa dependência, entre outros.

TAIANE SCHAEDLER PRASS - É mestre e doutora em matemática pela UFRGS e possui pós-doutorado em matemática pela UFRGS. É professora do Departamento de Estatística da UFRGS onde também atua no Programa de Pós-Graduação em Estatística como professor permanente. Possui uma gama diversificada de interesses de pesquisa, incluindo séries temporais, missing data e machine learning.

APLICAÇÕES DE MACHINE LEARNING:

Resultados de um Curso de Verão

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Ano 2024

APLICAÇÕES DE MACHINE LEARNING:

Resultados de um Curso de Verão

www.atenaeditora.com.br

contato@atenaeditora.com.br

@atenaeditora

www.facebook.com/atenaeditora.com.br

Ano 2024